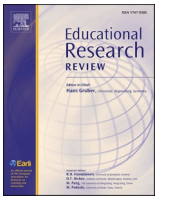
Educational Research Review 32 (2021) 100375

Contents lists available at ScienceDirect

Educational Research Review

journal homepage: www.elsevier.com/locate/edurev



Review

Is retell a valid measure of reading comprehension?

Yucheng Cao, Young-Suk Grace Kim *

University of California, Irvine, USA

ARTICLE INFO

Keywords:
Retell
Reading comprehension
Moderators
Assessment
Meta-analysis

ABSTRACT

Retell is used widely as a measure of reading comprehension. In this meta-analysis, we evaluated the relation between retell and other measures of reading comprehension among students in Grades 1–12. Data from 23 studies (82 effect sizes; $N = 4705$ participants) showed a moderate relation between retell and other measures of reading comprehension, $r = 0.46$. Moderation analyses revealed that the relation was stronger when reading comprehension was measured by cloze or maze tasks than when measured using a multiple-choice format. In addition, the relation was weaker in higher grades, but this was largely explained by text genre or the number of prompts in retell. The relation between 'oral' retell and reading comprehension was stronger with a greater number of prompts provided during retell tests. In contrast, results did not differ by other features of retell such as reading mode (oral or silent), text genres of retell (narrative or informational), or use of different oral retell evaluation methods (e.g., number of words or ideas, overall quality). Overall, the moderate magnitude of the relation between retell and other measures of reading comprehension indicates caution for using retell as the sole measure of reading comprehension. The results also indicate a need for a better understanding about more systematic approaches to retell assessment (e.g., number and kind of prompts in the case of oral retell) as a measure of reading comprehension.

1. Introduction

Reading comprehension is an integral skill for students' academic achievement and lifelong learning (National Institute of Child Health and Human Development, 2000). To monitor students' reading development and provide appropriate, timely instruction and interventions, accurate assessment of reading comprehension is essential. Reading comprehension involves complex information processing. According to the construction-integration model (Kintsch, 1988), comprehension processes include a meaning construction process, during which the reader constructs initial text-based representations using linguistic input of the given text, and an integration process, during which initial propositions are integrated into a coherent whole via inferential processes (also see other process-focused theoretical frameworks such as the constructionist models [e.g., Graesser, Singer, & Trabasso, 1994; Singer, Graesser, & Trabasso, 1994] and the landscape model [van den Broek, Rapp, & Kendeou, 2005]). Although details about comprehension processes differ across these models (see McNamara & Magliano, 2009 for an excellent review), all these models indicate that comprehension involves both automatic and effortful processes of construction, inference, and integration of meaning. Therefore, reading comprehension assessments should accurately capture these processes and associated mental representations of the text—mental representations of explicitly stated information *and* inferred information.

* Corresponding author. School of Education, University of California, Irvine, USA.
E-mail address: Youngsk7@uci.edu (Y.-S.G. Kim).

https://doi.org/10.1016/j.edurev.2020.100375
Received 15 November 2019; Received in revised form 7 October 2020; Accepted 2 December 2020
Available online 15 December 2020
1747-938X/

One widely used approach to assess students' reading comprehension is retell, particularly in the context of an informal or diagnostic assessment (Blachowicz & Ogle, 2008; Fuchs & Fuchs, 1992; Hansen, 1978; King, 1960; Kintsch, 1974; Kintsch & van Dijk, 1978; Matthew, 1996; Meyer & McConkie, 1973; Mills, Diehl, Birkmire, & Mou, 1993; Morrow, 1985, 1986; Sadoski, Goetz, & Fritz, 1993; Salvia & Hughes, 1990; Smith & Swinney, 1992; Stein & Glenn, 1975). In retell, students are typically asked and/or purposely prompted to tell or write down in their own words the main ideas or everything they can recall from a passage they just read (Bellinger & Diperna, 2011; Collins, Compton, Lindström, & Gilbert, 2020; Reed, Vaughn, & Petscher, 2012; Shapiro, Fritschmann, Thomas, Hughes, & McDougal, 2014; Tilstra & McMaster, 2013) or heard (Barnes, Kim, & Phillips, 2014; Kim & Schatschneider, 2017). It has been argued that retell taps into the interactive nature of reading comprehension, providing insight into students' comprehension processes and understanding of story structure; that is, retell manifests how students perceive, interpret, and organize the main idea units of the text they read (Kida, Ávila, & Capellini, 2016; Loyd & Steele, 1986; Matthew, 1996). For example, Loyd and Steele (1986) claimed that a modified retell method (paraphrastic recall that is evaluated by the reader's own perceptions rather than "expert judgments" of what is important about the text) could capture the reader-text interaction and permit "greater freedom of response than asking specific questions" (p. 2). By this account, retell is assumed to tap into both the lower and higher order comprehension processes. Retell is a free recall task where individuals are not prompted for inclusion of literal or inferred information, and therefore, the extent to which retell captures recall of given text (i.e., literal information or shallow comprehension) *and* higher order inferential processes (i.e., deep comprehension) would determine the relation between retell and other reading comprehension measures. Empirical evidence is limited, however.

Despite the wide use of retell, there has been insufficient research exploring the relation between retell and other measures of comprehension, and extant studies have shown inconsistent results. For example, some studies showed a moderate to strong relation (e.g., Fuchs, Fuchs, & Maxwell, 1988, $r = [0.59, 0.82]$; Marcotte & Hintze, 2009, $r = [0.46, 0.59]$), whereas others showed a weak relation (Loyd & Steele, 1986, $r = [0.26, 0.41]$; Reed et al., 2012, $r = [0.16, 0.26]$). These studies focused on different grade levels and involved a great variety of retell and reading comprehension assessments. Thus, our goal in the present study was to expand our understanding about the relation between retell and other measures of reading comprehension. We conducted a meta-analysis to determine the relation between retell and other measures of reading comprehension, and to explore the potential factors that may moderate the relation.

1.1. Potential factors that might influence the relation of retell to reading comprehension

A large body of studies has shown that child characteristics (e.g., word reading proficiency, language, and cognitive skills) contribute to one's reading comprehension (e.g., Florit & Cain, 2011; Kim, 2020). Another body of studies has indicated that assessment methods and text features explain variation in students' reading comprehension (e.g., Andreassen & Bråten, 2009; Collins et al., 2020; Eason, Goldberg, Young, Geist, & Cutting, 2012; Follmer & Sperling, 2018; García & Cain, 2014; Hall et al., 2014; McNamara, Ozuru, & Floyd, 2011; McNamara & Shapiro, 2005). The importance of these multiple factors (i.e., reader, assessment, and text features) in reading comprehension has been recognized in the RAND Reading Study Group (2002; also see Fox & Alexander, 2017). Furthermore, expanding the classic views such as the simple view of reading (Gough & Tunmer, 1986; Hoover & Gough, 1990) or the Reading Systems Framework (Perfetti & Stafura, 2014) which focus on reader/individual characteristics, recent theoretical models of reading explicitly state the roles of text features and assessments in reading comprehension as well as those of individual characteristics. For example, the Complete View of Reading (Francis, Kulesz, & Benoit, 2018) argues for incorporating text features such as passage-specific effects on one's performance in reading comprehension tasks. The direct and indirect effects model of reading (DIER; Kim, 2020) specifies interactions of text characteristics (including genre, decodability, language and cognitive demands) and assessment methods (e.g., retell, multiple-choice and open-ended formats; nature of questions) with individual characteristics in reading comprehension.

In line with these theoretical models and empirical evidence, in the present meta-analysis, we examined the following reader, assessment, and text factors that may moderate the relation between retell and reading comprehension: (1) the reading development phase of the individual (as operationalized by grade level), (2) features of reading comprehension assessment method (how children's reading comprehension is assessed), (3) features of retell assessment (whether children read the text orally or silently before the retell, whether retell is oral or written, how many times children are prompted during retell assessment, and in which genre retell is assessed), and (4) retell evaluation method (e.g., number of words, overall quality).

1.1.1. Reader characteristic: Reading development phase

Individual differences in reading proficiency determine the extent to which the reader understands what is read. Evidence indicates that the extent to which language and cognitive skills contribute to reading comprehension varies at different phases of reading development—in the beginning phase, decoding plays a large constraining role and dominates or largely determines one's reading comprehension skill whereas at a more advanced phase, comprehension skills are increasingly more important (Adlof, Catts, & Little, 2006; Florit & Cain, 2011; Hoover & Gough, 1990; Kim, 2020; Kim & Wagner, 2015). For example, even the very same test may measure different dimensions of comprehension as a function of children's reading proficiency (Cutting & Scarborough, 2006; Keenan, Betjemann, & Olson, 2008).

With regard to the relation between retell and other reading comprehension measures, Shapiro et al. (2014) found that retell had a slightly weaker relation with reading comprehension for fifth graders than for third graders. One potential explanation for this finding (grade as a proxy for developmental phase) may be that although students develop comprehension skills and associated language and cognitive skills, retell does not capture this fully in upper grades. For example, advanced readers are better at making inferences (Barth,

Barnes, Francis, Vaughn, & York, 2015; Cain & Oakhill, 1999; Yuill & Oakhill, 1991). If retell does not fully capture students' mental representation of inferred information and students in upper grades are better at making inferences, then the relation between retell and other reading comprehension measures will decrease in upper grades. In support of this, one study showed that retell (i.e., free recall) primarily captured shallow comprehension of texts for students between the ages of 11 and 15 years old (McNamara, Kintsch, Songer, & Kintsch, 1996). Another study showed that fourth-grade students' oral retell was primarily composed of recall of the given text with a limited amount of inferential information (Kim, Dore, Cho, Golinkoff, & Amendum, 2020).

Another explanation for a potentially differential relation between retell and other reading comprehension measures by development is that comprehension demands of texts change as a function of grade. Texts vary in many aspects including demands on language, content/topic knowledge, inference, and working memory (Kim, 2020; Kim & Petscher, 2020). By and large, texts are typically classified into narrative and informational texts. Narrative texts usually describe topics that are closely related to readers' personal life experiences and present causal relations (Williams, Hall, & Lauer, 2004). In contrast, informational texts usually introduce novel, unfamiliar concepts and information. Narrative and informational texts typically differ in many aspects including text structure, language demands, demands on content/topic knowledge, and density of information (Best, Floyd, & McNamara, 2008; Schleppegrell, 2001). Narrative texts are widely used in lower grades with increasing demands of and exposure to informational texts in upper grades (Graesser, McNamara, & Kulikowich, 2011). Therefore, if textual demands vary by genre and different genres of texts are used in different grades, the relation of retell and other measures of reading comprehension would change as a function of grade. Note that fully addressing the hypothesis about the role of text demands in the relation between retell and other reading comprehension measures requires accessing and coding texts. However, this was beyond the scope of this study as it was not possible to access the texts used in the primary studies. Instead, we examined whether the relation of retell and other reading comprehension measures varies as a function of grade, and if so, whether the moderation effect of grade remains after controlling for text characteristics operationalized as genre (see below).

1.1.2. Features of reading comprehension assessment method

Reading comprehension assessments vary in terms of method (e.g., multiple-choice questions, short open-ended questions, cloze task, and sentence verification or true/false tasks; Fuchs et al., 1988; García & Cain, 2014; Keenan et al., 2008), and the extent to which various tasks tap different comprehension processes may vary, thus influencing the relation between retell and other measures of reading comprehension. Multiple-choice format has been frequently used in standardized reading assessments as it requires less time and effort to administer with low teacher bias in scoring. Research has shown that multiple-choice tests are more predictive of student performance in comparison to in-class participation, written assignments, and case exams (Bontis, Hardie, & Serenko, 2009). Furthermore, when constructed well, multiple-choice items can tap into deep comprehension, such as inferential and evaluative comprehension, in addition to literal or shallow comprehension (see, for example, test construction in National Assessment of Educational Progress; Mazany, Pimentel, Orr, & Crovo, 2015). Similar to the multiple-choice format, short open-ended response format is also widely used and typically includes questions that are designed to tap into literal and inferential comprehension. The correlations between open-ended format and other formats (i.e., retell, cloze) have been shown to be moderate to high (Fuchs et al., 1988). However, unlike multiple-choice questions, short open-ended questions require a verbal response, which may put certain groups of students at a disadvantage, such as those with disabilities or English learners (Bellinger & Diperna, 2011; Collins et al., 2020).

Cloze tasks and maze tasks are also widely used to measure reading comprehension skill. In cloze tests, every fifth or seventh word is omitted from the passage, and students are required to restore the blanks with the words that were deleted. Similarly, a maze task requires students to select the word that best fits the meaning of the sentence from multiple options. A number of studies have shown that a maze task is sensitive to individual differences in reading achievement and indicative of student growth in reading performance, especially for students in upper elementary grades and above (Jenkins & Jewell, 1993; Shin, Deno, & Espin, 2000). On the other hand, studies also have indicated that cloze and maze tasks rely on vocabulary knowledge, syntactic understanding, and decoding skills (Howell & Nolet, 2000; Keenan et al., 2008), but not on higher order integration of information or inferential comprehension (Alderson, 1978; Francis, Fletcher, Catts, & Tomblin, 2005; Nation & Snowling, 1997; Parker, Hasbrouck, & Tindal, 1992; Suhorsky, 1975).

Finally, reading comprehension is also assessed using the sentence verification technique (SVT). In an SVT task, students are asked to read a passage and then answer each question (i.e., statement or sentence) by making a "yes" or "no" decision. For instance, if the stem statement has the same content (or different content) as the one in the passage, then the test takers are supposed to mark "yes" (or "no") next to it. Overall, the statements/sentences can be categorized into four types: (a) the original sentence extracted from the passage; (b) a paraphrased sentence, which replaces most of the words, but maintains the same meaning as the original sentence; (c) a meaning-altered sentence, which keeps most of the original sentence, but replaces just enough words to make the sentence differ in meaning from the original sentence; and (d) a new sentence that aligns with the passage theme and maintains similar structure (e.g., syntax), but does not relate to any passage sentence. In this way, proficient readers can be distinguished from poor readers by successfully identifying sentences of similar meaning with relatively less support from the linguistic structure (i.e., similar words or syntax; Marcotte & Hintze, 2009). One limitation of SVT is that deductive reasoning and inferential comprehension might be constrained to some extent (Cain & Oakhill, 2006). Nonetheless, Marcotte and Hintze (2009) revealed that SVT is strongly associated with the reading comprehension subtest of the Iowa Test of Basic Skills ($r = 0.73$).

In summary, reading comprehension is measured in various ways, and studies have shown that tasks vary in the extent to which they tap into different comprehension processes. Therefore, the relation of retell to other reading comprehension measures is likely to vary depending on reading comprehension assessment method. For instance, compared to multiple-choice or open-ended response formats (which typically include questions that are designed to elicit literal and inferential comprehension), cloze and maze tasks

might somewhat limited in capturing higher order inferential processes. In this way, the varying relation of retell to multiple-choice and open-ended tasks of reading comprehension versus to cloze and maze tasks of reading comprehension might indicate the extent to which retell captures or does not capture higher order inferential processes.

1.1.3. Features of retell assessment

Retell assessments can vary in several ways. First, retell assessments can vary in reading mode (oral or silent reading)—children are asked to read the text orally or silently before retelling. Reading mode may matter for the relation between retell and reading comprehension because oral and silent reading may differ in the extent to which they draw on cognitive processing (Hale et al., 2007). Reading aloud forces students to pay attention to each word whereas silent reading permits students to skip words, especially for beginning readers (Allington & McGill-Franzen, 2010; Hale et al., 2007; Schimmel & Ness, 2017). In fact, a recent eye-tracking study showed that first graders spent greater time looking at words during oral reading than during silent reading, and oral reading was more strongly related to their reading proficiency (Kim, Petscher, & Vorstius, 2019). On the other hand, oral versus silent reading for retell may have differential effects depending on reading proficiency. It is argued, based on the theory of limited cognitive resources, that the cognitive energy required in reading aloud (e.g., articulation) can be demanding for poor readers and hamper comprehension, while silent reading can free up more working memory for comprehension (Hale et al., 2011; Juel & Holmes, 1981; Rasinski, Rikli, & Johnston, 2009). Empirical evidence is somewhat mixed. For instance, some studies found that beginning readers and struggling readers tend to get better comprehension results when reading aloud (Fuchs et al., 1988; Kragler, 1995; S. D. Miller & Smith, 1989, 1985; Prior & Welling, 2001) whereas Reed and Petscher's (2012) study with students in Grades 6–8 showed that adding silent reading to oral reading did not significantly enhance retell performance. Then, the role of reading mode on the relation between retell and other reading comprehension measures may vary by grade.

Retell assessments can also vary by mode, oral or written retell (Bellinger & Diperna, 2011; Reed et al., 2012; Shapiro et al., 2014). Vieiro and García-Madruga (1997) found that written retell increased literal recall, while oral retell increased use of inferences and generalizations for third and fifth graders. Written retell might particularly strain student's performance for developing readers and writers because of the added challenges in writing or transcription skills (e.g., spelling). Overall, in their narrative review, Reed and Vaughn (2012) did not find a clear difference between oral and written retells. However, they did not conduct a statistical test for a potential difference as a function of oral versus written retell.

Additionally, retell assessments differ in terms of prompting conditions, which may influence the relation of retell to reading comprehension. Prompting conditions include the wording used to invite recall and the number of prompts. For example, Reed and Petscher (2012) found that the participants' retell scores were higher when they were prompted to tell "everything they remembered" in comparison to "tell in their own words what the passage was mostly about." Moreover, the students who received a follow-up prompt also had higher retell scores than those who only had the initial prompt (Reed & Petscher, 2012). Of course, this should not be taken simply as the greater the number of prompts, the higher the retell score because follow-up prompts may bring about incoherence if the student adds details that weaken the already presented coherent retell.

Another aspect of retell assessment includes text features. As noted above, texts vary in linguistic, cognitive, and knowledge demands (Kim, 2020; Kim & Petscher, 2020). Studies have shown that many children find narrative texts easier to understand, and children's comprehension of informational texts is weaker than their comprehension of narrative texts (Best et al., 2008; Graesser, McNamara, & Louwerse, 2003; Meyer & Ray, 2011; Mullis, Martin, Gonzalez, & Kennedy, 2003; Sáenz & Fuchs, 2002; Weaver & Kintsch, 1991; Williams et al., 2004), even after controlling for content (Wolfe & Mienko, 2007). Furthermore, a large amount of variation in children's performance in comprehension was attributed to differences among texts, and narrative versus informational genre explained all the differences in the variation among texts (Kim & Petscher, 2020). In retell, elementary grade students recalled less information from informational texts than from narrative texts (Best et al., 2008; Kim et al., 2020). If textual demands tend to differ by narrative and informational genres (Graesser, León, & Otero, 2002; Graesser et al., 1994; Trabasso & Magliano, 1996) and the amount of retell differs by genre (Best et al., 2008), it is plausible that what is captured in retell may vary by genre, which consequently would result in differential relations with other measures of reading comprehension.

Retell can also vary in terms of how it is evaluated or scored. Diverse scoring methods have been employed tapping different aspects of comprehension (Shapiro et al., 2014). Widely used evaluation methods include the number of total retold words, the number of idea units or propositions, the extent of story structure elements (e.g., goal, characters, setting), the use of certain linguistic features (e.g., coordinating conjunctions, subordinating conjunctions, mental and linguistic verbs), or the overall quality (assigning a single score after evaluating multiple aspects such as coherence, accuracy, and clarity). For example, in Dynamic Indicators of Basic Early Literacy Skills (DIBELS)-Retell Fluency subtest, a widely used assessment in U.S. schools, student retell is scored based on "the number of words the child retells that illustrate their understanding of the passage" (Good III & Good Kaminski, 2002, p. 37). Then, it is an open question whether these different approaches result in the same or different strengths of the relation between retell and other reading comprehension measures. For example, if evaluating overall quality captures multiple comprehension processes better than does counting the number of retold words, then it is reasonable to speculate differential relations as a function of retell evaluation method.

1.2. Present study

Despite the wide use of retell as a measure of reading comprehension, we do not have a good understanding about the relation between retell and other reading comprehension measures. Reed and Vaughn (2012) conducted a narrative review of retell as an indicator of reading comprehension. In the present study, we extend this work by conducting a comprehensive quantitative review of the studies and by examining moderators, using a meta-analytic approach. The following research questions guided the current

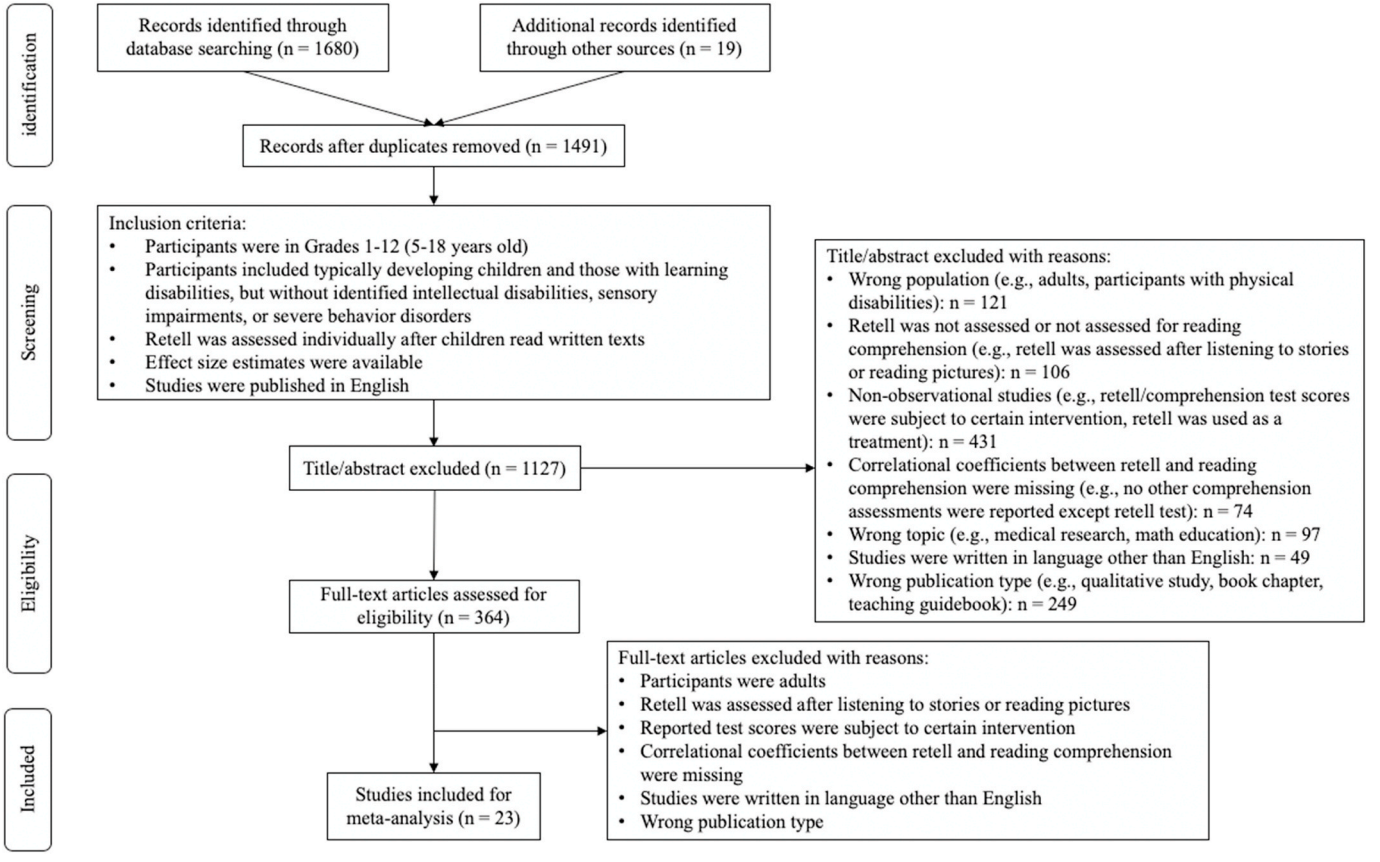


Fig. 1. PRISMA flow diagram showing the data screening process for the meta-analysis.

investigation:

(1) What is the relation of retell to other measures of reading comprehension for students in Grades 1–12?
(2) Does the relation vary as a function of grade level (a proxy for reading development phase), features of reading comprehension assessment method (e.g., multiple-choice, open-ended response, cloze task), features of retell assessment method (reading mode, retell mode, number of prompts, and narrative and informational genre), and retell evaluation method?
(3) If there is a moderation effect of grade, is it explained by the other moderators (i.e., reading comprehension assessment method, retell assessment method, or retell evaluation method)?
(4) Does the effect of reading mode vary depending on grade level?

We expected a positive, moderate relation between retell and other reading comprehension measures based on a narrative review by Reed and Vaughn (2012). We also hypothesized that the relation might vary as a function of reading comprehension assessment method to the extent that different methods measure integration and inference of information to a different degree. If retell largely captures recall of the provided text rather than higher order inferential processes, then its relation to maze and cloze tasks might be stronger than to multiple-choice tasks because previous studies suggested that maze and cloze tasks tended to tap into lower order comprehension processes (i.e., literal comprehension) whereas multiple-choice tasks tended to tap into lower and higher order inferential processes when designed well (i.e., deep comprehension; Mazany et al., 2015; McNamara et al., 1996). Likewise, we expected that retell evaluation methods such as overall quality or story structural elements might capture multiple comprehension processes and thus, we posited a stronger relation of retell with comprehension when retell is evaluated for overall quality and story structural elements than simple counts of words. In addition, we expected that features of retell assessment method such as the number of prompts might positively influence the magnitude of the relation between retell and reading comprehension (Reed & Petscher, 2012). However, we did not have clear hypotheses for the other moderators of retell assessment method (e.g., oral vs. silent reading mode). We hypothesized that if there is a moderation effect of grade level, this could be explained by other moderators such as reading comprehension assessment method and retell assessment method (e.g., genre) because these features might vary across grades.

2. Method

2.1. Search procedure

The main literature search was conducted using three electronic databases, ERIC, PsycINFO, and Sociological Abstracts, through ProQuest and two additional electronic databases, Academic Search Complete and Education Source, through EBSCO. The search was restricted to retell–reading comprehension studies of students in Grades 1 to 12 using the following combination of terms: all(retell* OR recall) AND all(read* comprehen*) AND all("grade 1″ OR "grade 2″ OR "grade 3″ OR "grade 4″ OR "grade 5″ OR "grade 6″ OR "grade 7″ OR "grade 8″ OR "grade 9″ OR "grade 10″ OR "grade 11″ OR "grade 12″ OR "primary education" OR "secondary education" OR "elementary school" OR "middle school" OR "junior high school" OR "high school"). In addition, a secondary literature search was conducted by manually reviewing the reference list of a prior pertinent narrative synthesis (i.e., Reed & Vaughn, 2012). The above electronic and manual searches covered all of the papers that were available through 2017. We obtained a total of 1699 papers for possible inclusion in the meta-analysis.

2.2. Inclusion and exclusion criteria

By the end of January 2018, all the identified records were imported to Covidence systematic review software (Veritas Health Innovation, 2018) for initial title and abstract screening. Of all the collected 1491 documents (after duplicates were removed), 1127 were excluded by title/abstract screening (see Fig. 1) and the eligibility of the remaining articles was assessed by full-text screening by two raters independently (raters have expertise in literature search and literacy education).

To ensure that the included articles were most relevant to our research topic and questions, inclusion criteria were selected as follows: (1) participants were in Grades 1–12 (5–18 years old); (2) participants included typically developing children regardless of their English learner status and those with learning disabilities (dyslexia, language impairment, etc.), but without identified intellectual disabilities, sensory impairments, or severe behavior disorders; (3) retell was assessed individually after children read written texts (4) studies reported either effect size estimates (e.g., correlation coefficients); or data for determining effect estimates between retell and other measures of reading comprehension (standardized and experimental measures); and (5) studies were published in English. Dissertations and other unpublished papers as well as peer reviewed articles were included in order to reduce potential publication bias.

In title/abstract screening, studies were excluded because reported retell and other comprehension test scores were subject to a certain intervention (38%; if the relation between retell and other comprehension scores was reported at pretest, those studies were included); they were not quantitative studies (e.g., teaching guidebook, book chapter, qualitative study; 22%); retell or other comprehension measures were not assessed (16%); participants were adults or with physical disabilities (11%); studies were irrelevant to reading comprehension (e.g., medical research, math education; 9%); or studies were published in languages other than English (4%). In full-text screening, studies were excluded mainly for not providing information on the correlations between retell and other reading comprehension assessment results (see Fig. 1). All disagreements were resolved through discussion (interrater reliability = 98%). Among the 33 potential studies to be included, 13 studies did not report necessary effect sizes, and these studies' authors were

contacted. Of the 13 authors contacted, three authors provided data, and therefore, a total of 23 studies (82 effect sizes) was used for the current analysis ($N = 4705$ participants), of which 21 are journal articles and 2 are doctoral dissertations.

2.3. Coding procedures

The coding protocol was developed in an iterative manner. The preliminary coding form was designed to capture main study information including participants' description, reading assessment types and scores, and other general study features. A random subsample of the included studies (n = 10) were coded by two trained researchers (the first author and a doctoral student both of whom screened the articles) independently after reaching agreement on the interpretation of each coding category. The interrater reliability ranged from 0.84 to 1.00 with an average of 92% for the initial coding procedure. Subsequently, the coders modified the coding scheme while resolving all disagreements. For the main and moderator analyses, each study was coded in the following three categories: a) sample characteristics, b) assessment features (reading comprehension assessment, retell assessment, and retell evaluation methods; see above), and c) effect sizes.

2.3.1. Sample characteristics

We coded the sample characteristics including the study location, predominant ethnicity, primary language, English language learner status, socioeconomic status, sample size, gender/sex, grade level, age, and learning disability status (see Table 1).

2.3.2. Assessment features

Reading comprehension assessment features were coded for the moderator analysis (i.e., multiple-choice, short open-ended, cloze or maze, sentence verification, or mixed methods). Retell assessment features were coded for reading mode (i.e., oral or silent), retell mode (i.e., oral or written), number of prompts, and text genre (i.e., narrative or informational). Retell evaluation method was coded for word counts, propositions or idea units, story structure elements, or overall quality (see Table 2).

Table 1
Sample characteristics of individual studies.

Study author(s)	Primary ethnicity	Primary language	% ELL status	SES	Sample size	% Boy	% Girl	Grade	Age	% LD status
Applegate et al. (2009)	Caucasian	N/A	N/A	N/A	171	35	65	2 to 10	N/A	N/A
Atkins and Cummings (2011)	N/A	N/A	N/A	N/A	104	N/A	N/A	3 & 4	N/A	N/A
Bellinger and Diperna (2011)	Caucasian	English	0	Diverse	44	47.7	52.3	4	[9,10]	N/A
Carlisle (1999)	Caucasian & African American	N/A	N/A	Diverse	63	N/A	N/A	6 & 8	N/A	30.2
Carroll (1985)	Diverse	English	0	Diverse	30	N/A	N/A	5	N/A	N/A
Chaka and Booi-Ncetani (2015)	African	ELL	100	N/A	17	59	41	10	N/A	N/A
Hansen (1978)	Diverse	N/A	N/A	Diverse	34	73.5	26.5	5 & 6	N/A	50
Keenan et al. (2008)	N/A	English	0	N/A	510	N/A	N/A	N/A	[8,18]	0
Loyd and Steele (1986)	N/A	N/A	N/A	N/A	108	N/A	N/A	11 & 12	N/A	0
Marcotte and Hintze (2009)	Caucasian & Hispanic	English	30	Low	111	50	50	4	N/A	15
McNamara et al. (2011)	Caucasian & African American	English	3.08	Diverse	65	47.7	52.3	4	[9,11]	N/A
Pflaum (1980)	N/A	N/A	N/A	Low	76	61.8	38.2	4 to 7	[9,12]	47.4
Rasinski (1984)	N/A	N/A	N/A	Mid	142	N/A	N/A	3 & 5	N/A	0
Reed and Petscher (2012)	Hispanic & Caucasian	N/A	8	Low & Mid	527	46	54	6, 7, & 8	N/A	5
Reed et al. (2012)	Hispanic & African American	N/A	24	Low	311	50	50	7 & 8	N/A	34
Riedel (2007)	African American	English	3.9	Low	1518	50	50	1	N/A	0
Roberts et al. (2005)	African American	N/A	N/A	Low	86	N/A	N/A	1	N/A	N/A
Schimmel and Ness (2017)	Caucasian	English	0	N/A	48	47.9	52.1	4	[9,10]	N/A
Shapiro et al. (2014)	Caucasian	N/A	N/A	Diverse	271	N/A	N/A	3 & 5	N/A	0
Shinn et al. (1992)	Caucasian	N/A	N/A	N/A	238	51	49	3 & 5	N/A	5
Thomas (2012)	Caucasian	N/A	N/A	N/A	107	52.3	47.6	3	[8,9]	6.5
Walczyk and Raska (1992)	Caucasian	N/A	N/A	Mid	124	47.6	52.4	2, 4, & 6	N/A	N/A

Note. N/A = not available (primary studies did not report information); ELL = English Language Learners; SES = socioeconomic status; LD = learning disabilities. All the studies were conducted in the U.S. except Chaka and Booi-Ncetani (2015), which was conducted in South Africa.

Table 2
Information on coded moderators.

Study author(s)	Features of reading comprehension assessment method	Features of retell assessment				Retell evaluation method
		Reading mode	Retell mode	# of Prompts	Genre	
Applegate et al. (2009)	Open ended[s]	Both	Oral	2	Narrative	SSE
Atkins and Cummings (2011)	Mixed (multiple choice, open-ended)[s]	Oral	Oral	2	N/A	Word counts
Bellinger and Diperna (2011)	Cloze and vocabulary[s]	Oral	Oral	2	N/A	Word counts
Carlisle (1999)	SVT[e]	N/A	Oral	2	Informational	Propositions/Idea units
Carroll (1985)	Mixed (multiple choice, short answer, etc.)[s]	N/A	Oral	2	Narrative	SSE and quality
Chaka and Booi-Ncetani (2015)	N/A	N/A	Oral	N/A	N/A	N/A
Hansen (1978)	Open ended[e]	Oral	Oral	2	Narrative	Propositions/Idea units
Keenan et al. (2008)	Short answer, cloze and multiple choice[s]	Oral	Oral	N/A	Both	Propositions/Idea units
Loyd and Steele (1986)	Multiple choice[s]	N/A	Written	N/A	N/A	Gist and quality
Marcotte and Hintze (2009)	Maze, multiple choice, SVT[s]	Both	Both	2 & 5	N/A	Word counts
McNamara et al. (2011)	Multiple choice[s]	Silent	Oral	1 & 3	Both	Propositions/Idea units
Pflaum (1980)	Multiple choice[s]	Oral	Oral	3	N/A	Propositions/Idea units
Rasinski (1984)	Multiple choice[e&s]	Oral	Oral	N/A	Informational	Propositions/Idea units
Reed and Petscher (2012)	Multiple choice[s]	Both	Oral	1 & 2	Both	Propositions/Idea units
Reed et al. (2012)	Multiple choice, clo/maze[s]	Oral	Oral	2	Both	Main ideas and quality
Riedel (2007)	Multiple choice[s]	Oral	Oral	2	N/A	Word counts
Roberts et al. (2005)	Cloze and vocabulary[s]	Oral	Oral	N/A	N/A	Relevant words
Schimmel and Ness (2017)	Open ended[s]	Both	Oral	N/A	Both	Propositions/Idea units
Shapiro et al. (2014)	Mixed (multiple choice, open-ended)[s]	Oral	Oral	N/A	Narrative	SSE and word counts
Shinn et al. (1992)	Multiple choice and cloze[e&s]	Silent	Written	4	Narrative	Recognizable words
Thomas (2012)	Multiple choice, Mixed (multiple choice, open-ended)[e&s]	Oral	Oral	2 & 3	Both	SSE and word counts
Walczyk and Raska (1992)	Multiple choice[e&s]	Silent	Oral	1	Narrative	Inferences

Note. N/A = not available (primary studies did not report information); SSE = story structure elements; SVT = sentence verification technique; clo/maze = cloze and maze. Reading assessment method with letter superscript *s* refers to standardized measures and with letter superscript *e* refers to experimental measures.

2.3.3. Effect sizes

The correlation coefficients between retell and other reading comprehension measures were coded (see forest plot in Fig. 2). The two coders applied the finalized coding categories to an additional random subsample (n = 10) of the included studies. Interrater reliability ranged from 95% for the assessment and text characteristics to 100% for sample characteristics and effect sizes.

2.4. Meta-analytic procedures

2.4.1. Effect size estimates

All calculations and statistical analyses were conducted in the RStudio open-source software (Version 3.6.3, R Core Team, 2018; Version 1.2.5033, RStudio Team, 2016) using functions available in the metafor package (Version 2.0-0; Viechtbauer, 2010) and the Robumeta package (Version 2.0; Fisher, Tipton, & Zhipeng, 2017). The effect size index used in the current study was the correlation coefficient, *r*. We interpreted the magnitude of the correlation as suggested by Cohen (1992): $r = 0.10$ is considered small, $r = 0.30$ is moderate, and $r = 0.50$ or larger is large. To ensure the sampling distributions of *r* values are normal, we applied Fisher's *z* transformation to our data analysis and reconverted *z* values back to *r* values for the presentation and interpretation of results (Borenstein, Hedges, Higgins, & Rothstein, 2011).

2.4.2. Statistical analysis

Given that the included studies were conducted independently by diverse researchers with a variety of participants and measures, we employed a random-effects inference model (Borenstein et al., 2011). In contrast to a fixed-effects model, this model treats the

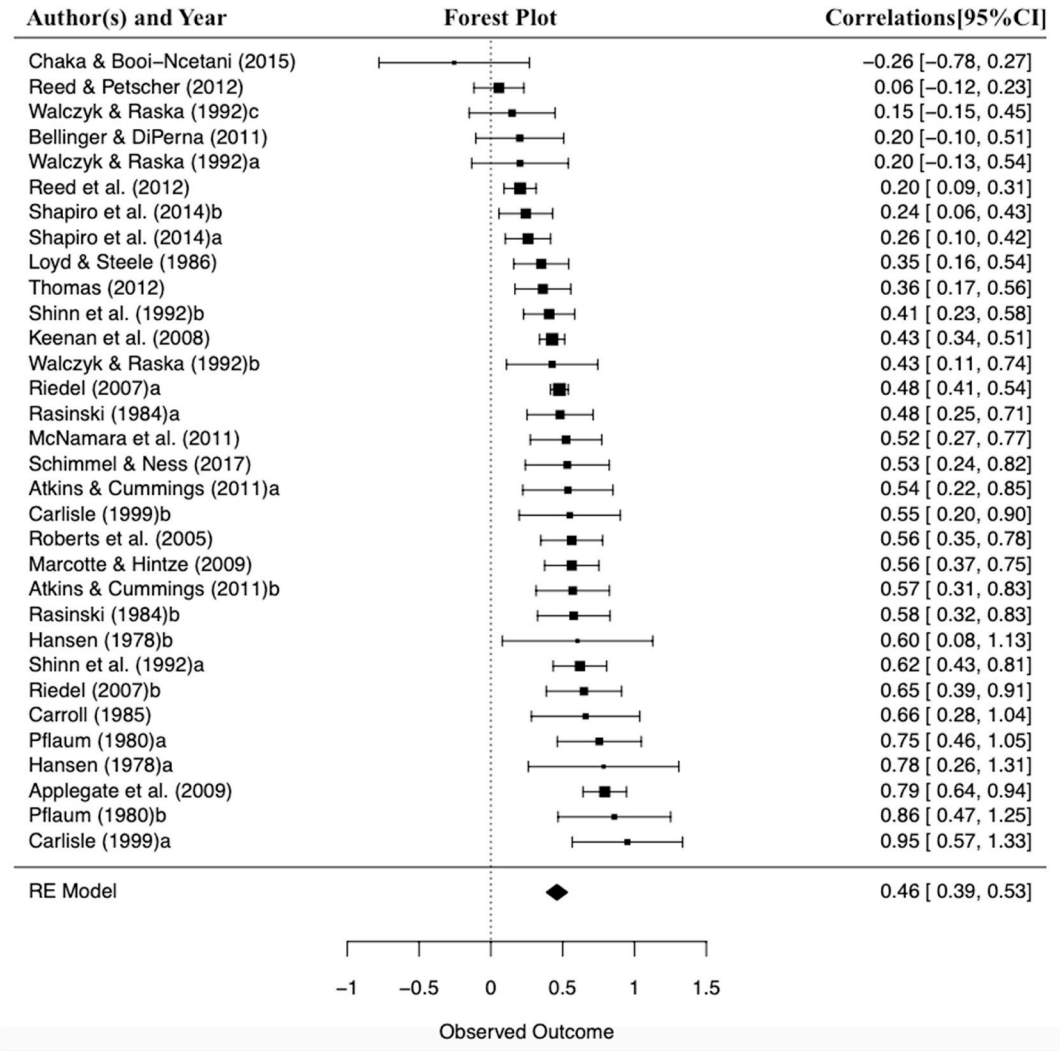


Fig. 2. Forest plot with 95% confidence interval and estimated mean correlation for each study with an overall weighted mean correlation at the bottom. Studies with alphabetic endings refer to the multiple independent correlations generated from the same study. RE model = Random Effects model.

studies as samples from a greater population and takes between-study variation into consideration by adding a random term in the model (Nakagawa & Santos, 2012; Viechtbauer, 2010), thus allowing us to meet the study goal of generalizing results beyond the included studies. We employed maximum likelihood estimation as it is closer, on average, to the true parameter values with a lower sampling variability in comparison to other alternative estimators (Vevea & Coburn, 2015).

Our data structure is complex as many studies included subgroups of participants and reported multiple effect sizes for each subgroup. To handle statistically dependent effect sizes, we employed robust variance estimation (RVE), which accounts for both hierarchical effects cases (i.e., multiple studies are nested within a larger cluster) and correlated effects cases (i.e., the same participant group provides multiple effect sizes) with small-sample corrections (Hedges, Tipton, & Johnson, 2010). However, RVE neither models heterogeneity at multiple levels nor provides corresponding hypothesis tests, and the power of the categorical moderator highly depends on the features of the covariate (Tanner-Smith, Tipton, & Polanin, 2016). When the number of studies is small, the test statistics and confidence intervals based on RVE can have inflated Type I error (Hedges et al., 2010; Tipton & Pustejovsky, 2015). Consequently, we used RVE mainly for the main analysis of the overall weighted mean correlation. Although we performed moderation analyses using RVE, these results should be taken with caution as RVE requires a sufficiently large sample size and may yield inaccurate *p* values when degrees of freedom are under 4 (which occurred in some moderators of this study; see Tables 5 and 6). Given the limitations of RVE, we also fit our data into multilevel/multivariate random-effects models (one level at subgroup and one level at study) for both the main and moderation analyses using the function rma.mv in metafor. Although this approach is more appropriate for hierarchical effects cases, it could meet the goals of operating heterogeneity, moderation, and sensitivity analyses which are not currently available for RVE.

Table 3
Multilevel random effects model: Meta-regression of moderators on overall weighted average correlation (Fisher's z).

Moderator	*k*	Intercept	*b*	95% CI	*Qm(df)*
Grade level	82	.62***(.08)	-.04*(.01)	[-.06, −.01]	6.57*(1)
Reading comprehension assessment method	81				7.11(4)
Multiple-choice[1]	39	.44***(.05)		[.35, .53]	
Short open-ended	7		.08(.05)	[-.02, .19]	
Cloze/maze	11		.08*(.03)	[.01, .15]	
Sentence verification	4		.02(.08)	[-.13, .18]	
Mixed	20		-.04(.05)	[-.15, .07]	
Reading mode in retell	73				1.88(1)
Oral[1]	56	.40***(.05)		[.30, .50]	
Silent	17		.09(.07)	[-.04, .22]	
Number of prompts in oral retell	53	.25*(.10)	.11**(.04)	[.03, .19]	7.77**(1)
Text genre in retell	58				2.84(2)
Narrative[1]	32	.51***(.07)		[.38, .64]	
Informational	12		-.03(.05)	[-.13, .08]	
Both	14		-.18(.11)	[-.40, .03]	
Evaluation method in oral retell	70				.73(3)
Word counts[1]	26	.41***(.08)		[.26, .56]	
Propositions/idea units	28		.06(.10)	[-.14, .27]	
Story structure elements	10		.01(.04)	[-.07, .10]	
Mixed	6		.10(.15)	[-.18, .39]	

Note. Moderator with numeric superscript 1 refers to the reference group in the independent moderation analysis. Standard error in parentheses in Intercept and *b* columns. k = number of effect sizes included in analysis; Q_m = test of difference on moderators. The number of effect sizes in each grade level: Grade 1 = 5, Grade 2 = 4, Grade 3 = 22, Grade 4 = 16, Grade 5 = 16, Grade 6 = 7, Grade 7 = 9, Grade 8 = 1, Grade 11 = 2. *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

Table 4
Multilevel random effects model: Meta-regression of moderators controlling for grade level.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Grade level	-.03*(.01)	-.04**(.02)	-.04(.02)	-.02(.02)	-.04*(.02)	-.04*(.02)
Reading comprehension assessment method						
Short open-ended	.09(.05)					
Cloze/maze	.08*(.03)					
Sentence verification	.04(.08)					
Mixed	-.05(.05)					
Reading mode in retell						
Silent		.09(.06)				.18(.16)
Number of prompts in oral retell			.11**(.04)			
Text genre in retell						
Informational				-.03(.05)		
Both				-.16(.11)		
Evaluation method in oral retell						
Propositions/idea units					.14(.10)	
Story structure elements					.01(.04)	
Mixed					.21(.14)	
Grade*Reading mode						-.02(.04)
Intercept	.58***(.08)	.58***(.08)	.41**(.14)	.62***(.12)	.53***(.09)	.55***(.08)
K	81	73	53	58	70	73

Note. Reference group for each moderator was omitted: multiple-choice in reading assessment, oral reading in reading mode, narrative in genre, word counts in retell evaluation. Standard error in parentheses. k = number of effect sizes included in analysis. *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

2.4.3. Heterogeneity and moderator analysis

The homogeneity statistic Q was computed to assess the variation in correlations between studies (Lipsey & Wilson, 2001). A significant and large Q value suggests that the differences among effect sizes, that is, the variability among the correlations within the data set, cannot be explained by sampling error alone. The Q-test is usually supplemented by the I^2 statistic for quantifying the percentage of variation across studies due to real heterogeneity. By convention, I^2 values range from 0% to 100% and indicate higher level of heterogeneity with increasing percentage: small = 25%, moderate = 50%, and high = 75% (Higgins, Thompson, Deeks, & Altman, 2003). When significant heterogeneity was confirmed, moderation analyses were conducted to explain the between-study differences and to identify potential effects of moderating factors. Concerning the inconsistent missing data in moderators, we first investigated each moderator independently, then followed this up with models where grade level was controlled as a covariate in each meta-regression model.

Table 5
RVE: Meta-regression of moderators on overall weighted average correlation (Fisher's z).

Moderator	*k*	Intercept	*b*	95% CI
Grade level	82	.59***(.07)	-.03(.02)	[-.07, .01]
Reading comprehension assessment method	81			
Multiple-choice[1]	39	.44***(.05)		[.32, .55]
Short open-ended	7		.22(.10)	[-.03, .47]
Cloze/maze	11		.00(.09)	[-.21, .21]
Sentence verification	4		.26(.16)	[-.35, .86]
Mixed	20		-.03(.09)	[-.23, .18]
Reading mode in retell	73			
Oral[1]	56	.44***(.05)		[.33, .54]
Silent	17		-.01(.08)	[-.20, .18]
Number of prompts in oral retell	53	.06(.15)	.22*(.07)	[.06, .39]
Text genre in retell	58			
Narrative[1]	32	.45***(.07)		[.29, .60]
Informational	12		.13(.11)	[-.13, .39]
Both	14		-.11(.11)	[-.38, .15]
Evaluation method in oral retell	70			
Word counts[1]	26	.46***(.04)		[.36, .56]
Propositions/idea units	28		.03(.08)	[-.14, .20]
Story structure elements	10		-.17(.07)	[-.40, .05]
Mixed	6		.07(.22)	[-.58, .71]

Note. Moderator with numeric superscript 1 refers to the reference group in the independent moderation analysis. Standard error in parentheses. k = number of effect sizes included in analysis. *$p < 0.05$. ***$p < 0.001$.

Table 6
RVE: Meta-regression of moderators controlling for grade level.

	Model 1	Model 2	Model 3	Model 4	Model 5
Grade level	-.03(.01)	-.05(.02)	-.02(.03)	-.00(.03)	-.04(.04)
Reading comprehension assessment method					
Short open-ended	.23(.10)				
Cloze/maze	-.02(.08)				
Sentence verification	.31(.16)				
Mixed	-.05(.09)				
Reading mode in retell					
Silent		-.00(.08)			
Number of prompts in oral retell			.22*(.07)		
Text genre in retell					
Informational				.14(.12)	
Both				-.11(.12)	
Evaluation method in oral retell					
Propositions/idea units					.12(.11)
Story structure elements					-.14(.06)
Mixed					. 20(.21)
Intercept	.57***(.07)	.62***(.08)	.13(.14)	.47*(.15)	.58***(.10)
K	81	73	53	58	70

Note. Reference group for each moderator was omitted: multiple-choice in reading assessment, oral reading in reading mode, narrative in genre, word counts in retell evaluation. Standard error in parentheses. k = number of effect sizes included in analysis. *$p < 0.05$. ***$p < 0.001$.

2.5. Sensitivity analyses

2.5.1. Influential outliers

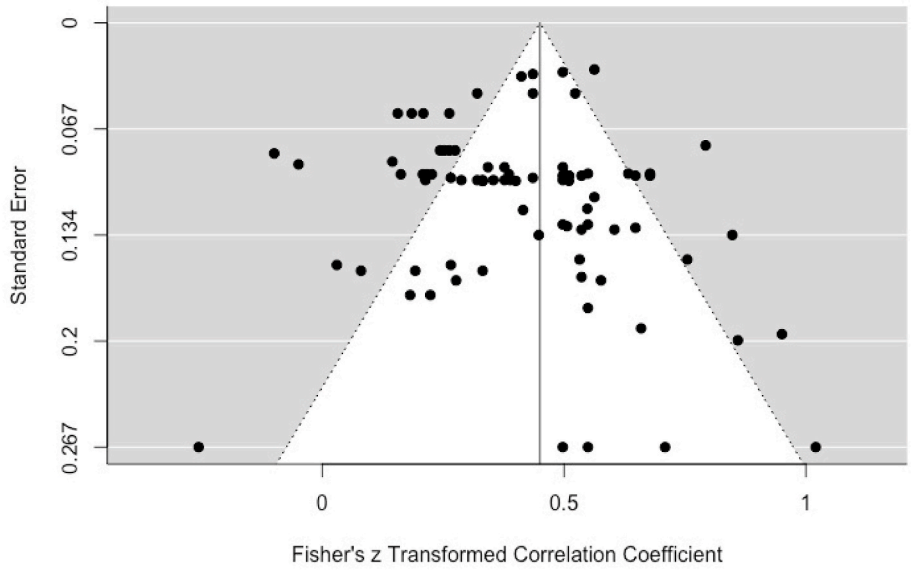
Diagnostic tests for identifying outliers and influential cases together with the comparison between the fitted models with and without outliers/influential cases were performed in an effort to evaluate the sensitivity of our analyses. By convention, we defined influential outliers as effect sizes with hat values exceeding twice the average hat value (which is 0.024 in our dataset) and with Cook's distance estimates greater than 0.048 (cutoff = 4/number of observations). In general, six effect sizes were identified as outliers (Applegate, Applegate, & Modla, 2009; Carroll, 1985; Chaka & Booi-Ncetani, 2015; Pflaum, 1980; Roberts, Good, & Corcoran, 2005; Schimmel & Ness, 2017; see Appendix A). Sensitivity analysis showed that, after removing them, the overall correlation decreased to $r = 0.42$ ($SE = 0.04$, 95% CI [0.34, 0.51], $p < .001$) from $r = .46$ (see the Results section) and the significance level remained statistically and positively significant. This result, therefore, confirmed the moderate and significant correlation (see Results section below) between retell and reading comprehension.

2.5.2. Publication bias

Studies with statistically significant results are more likely to be published. Hence, including only published studies for meta-

analysis might lead to an overestimated mean effect size (MES). To determine the presence of publication bias in our analysis, we performed a funnel plot and Egger's regression test (Sterne & Egger, 2006). The funnel plot indicated that studies are missing to the left of the mean near the bottom with several small studies showing larger effect sizes without smaller effect size counterparts (see Fig. 3:

a)



b)

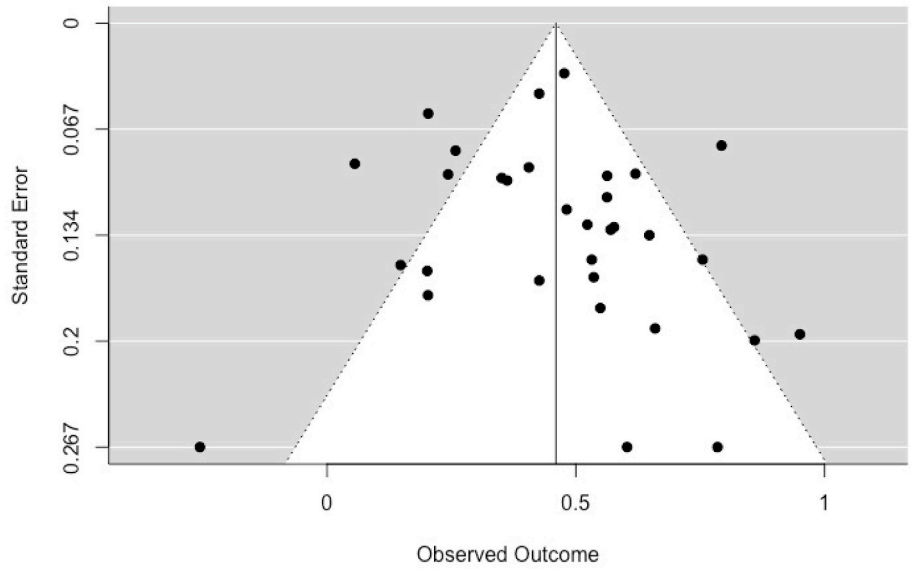


Fig. 3. Funnel plot representing correlations generated by separate cases ($k = 82$ in Fig. 3a; and $k = 32$ in Fig. 3b) nested within independent studies (n = 22) against corresponding standard error.

3a displays all the 82 effect sizes, and 3b displays the 32 effect sizes, which are the average values by grade level for each study; Fig. 4 also shows funnel plot with filled-in data). For the Egger's regression test, we modified the original multilevel/multivariate random-effects model by adding the standard error of the effect sizes as a covariate to further determine whether the relation between the precision and size of the sampled studies is as asymmetrical as the funnel plot demonstrated. The intercept significantly deviated from 0 ($b = 0.34, p < .001$). As the intercept is expected to be zero in the absence of publication bias, our result suggested the existence of a publication bias.

To determine whether the publication bias operates mainly on smaller studies, we performed cumulative meta-analysis by restricting our analysis to larger studies for robustness. As shown in Appendix B, the studies were sorted by size/precision, and effect sizes in the first three rows matched the asymmetric distribution near the bottom of the funnel plot. Accordingly, we removed two studies (Chaka & Booi-Ncetani, 2015; Hansen, 1978) and re-estimated the overall correlation, $r = 0.45$ ($SE = 0.04$, 95% CI [0.37, 0.54], $p < .001$), which was not different from the estimate including these studies (see the Results section below). The result indicated that including small studies did not impact our analyses.

2.5.3. Study quality and standardized versus experimental reading comprehension measures

The quality of studies was evaluated using the Quality Assessment Tool for Observational Studies (National Heart, Lung, and Blood Institute, n. d.). We adapted the original version to assess the study quality mainly in terms of its research question, study population, reading assessment, outcome report, and statistical analysis (see Appendix C). The overall quality was rated as "strong," "moderate," or "weak." Overall, 13 investigations were rated as strong, seven were moderate, and two were weak. Considering that meta-analytical studies might be vulnerable to low-quality studies, we excluded the two studies that were rated as weak in sensitivity analysis (i.e., Chaka & Booi-Ncetani, 2015; Roberts et al., 2005). However, the sensitivity analysis indicated that study quality has little impact on the analyses, because excluding the two studies did not change the significance level or magnitude of the overall correlation, $r = 0.46$, $SE = 0.05$, 95% CI [0.37, 0.55], $p < .001$. Hence, including low-quality studies did not influence our findings.

We also examined whether the nature of reading assessments, standardized versus experimental measures, is a confounder in the relation between retell and reading comprehension. We split the effect sizes into two groups based on the standardized versus experimental nature of assessment. The comparison between these two groups of studies yielded non-significant results, suggesting that the assessment type does not influence our analyses ($Q_m = 1.52$, $df = 1$, $p = .22$).

3. Results

3.1. Descriptive information

An overview of the 23 studies (82 effect sizes; $N = 4705$ participants), representing 22 separate investigations (two studies were treated as one since they used the same sample and generated the exact same results published in versions of journal article and dissertation respectively), with sample characteristics and coded moderators is displayed in Tables 1 and 2, respectively. As shown in

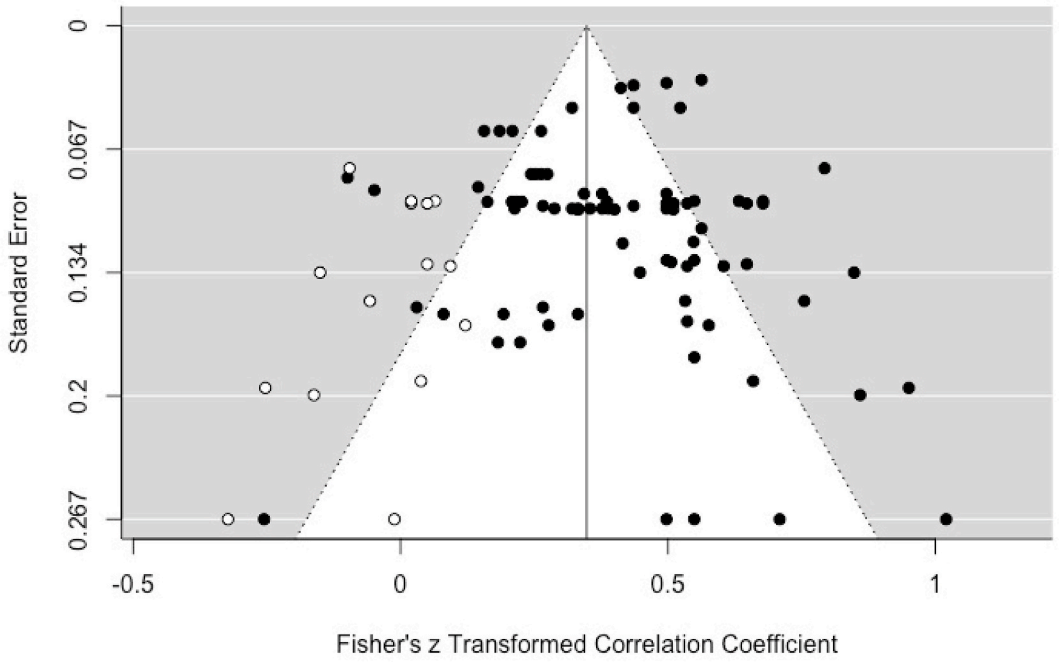


Fig. 4. Funnel plot with filled-in data based on the trim-and-fill method.

Table 1, all but one of these studies were conducted in the United States (the exception was in South Africa). Of the 17 studies providing information on ethnicity, seven involved primarily Caucasians, whereas six studies employed diverse populations and four studies focused mainly on minority students (i.e., Hispanics and African Americans). Only four studies reported recruiting exclusively English speakers, and six studies specified English language learners in the sample. Of the 14 studies indicating participant SES status, six studies contained students of low-SES background, six studies included students from diverse backgrounds, and two included students mainly from the middle-class. Sample sizes varied, ranging from 17 to more than 1000. The proportion of boys to girls was comparable in many studies. In addition, more than one half of the studies included students from Grades 3 to 5 (n = 12), and eight studies involved students with learning disabilities.

Table 2 shows that many studies employed multiple-choice items in reading comprehension assessments (n = 15), followed by short open-ended questions (n = 8), cloze and maze (n = 6), and sentence verification (n = 2). According to the studies reporting reading mode (oral or silent reading) and text genre of retell (narrative or informational), almost half of them required students to read orally before retelling (n = 11) and used narrative passages (n = 6). The majority of studies used oral retell assessment, two studies examined written retell performance, and one study administered both. Retell protocols were usually scored by word counts (n = 8), number of propositions or ideas (n = 8), story structure elements (n = 4), and overall quality (n = 3). Also note that many studies did not provide full information on the reliability and validity of retell and other reading comprehension assessments. Of those which reported, reliability ranged from 0.57 to 1.00, and validity ranged from 0.39 to > 0.90 (see Appendix D).

3.2. Research question 1: What is the relation of retell to other measures of reading comprehension for students in grades 1–12?

The overall magnitude of the relation between retell and reading comprehension was significant and moderate (see Fig. 2). In the RVE model, an average effect was found: $r = 0.46$, 95% CI [0.38, 0.54], $z = 11.9$, $p < .001$. The multilevel/multivariate random-effects model yielded a similar result, $r = 0.45$, 95% CI [0.36, 0.54], $z = 10.08$, $p < .001$. The heterogeneity test was statistically significant, $Q = 357.66$, $df = 81$, $p < .001$, and the I^2 value indicated that 84.86% of the total observed variance was due to between-study differences rather than within-study sampling error. Accordingly, additional moderator analyses were performed in an effort to explain the substantial heterogeneity among the effect sizes.

3.3. Research question 2: Does the relation vary as a function of grade level (a proxy for reading development phase), features of reading comprehension assessment, retell assessment features (reading mode, retell mode, number of prompts, text genre), and retell evaluation method?

Results of moderation analyses are presented in Table 3. Each moderator was evaluated individually. The reference group for each categorical variable was the subcategory that had the most effect sizes and was predominantly employed in the included studies. Specifically, the following were used as the reference groups: multiple-choice task, oral reading mode, narrative text in retell, and word counts for retell evaluation, respectively.

3.3.1. Student grade level

For the studies that provided an aggregated effect size for participants of multiple grade levels, we took the average of the grade levels or the weighted average grade based on the reported sample size in each grade level. Analysis of average grade level as a continuous variable demonstrated a significant negative moderating effect, $Q_m = 6.57$, $df = 1$, $p < .05$; $b = -0.04$, $SE = 0.01$, $p < .05$ (see Table 3), such that the average correlation between retell and reading comprehension was weaker in higher grade levels (−0.04 per grade).

3.3.2. Features of reading comprehension assessment method

The omnibus test was nonsignificant, $Q_m = 7.11$, $df = 4$, $p = .13$ (see Table 3). However, on average, studies using cloze or maze yielded a significantly stronger association, $b = 0.08$, $SE = 0.03$, $p < .05$, than those using multiple-choice format (0.44). By contrast, the MES obtained from studies using multiple-choice items was not significantly different from those using short open-ended format, $b = 0.08$, $SE = 0.05$, $p = .11$; sentence verification, $b = 0.02$, $SE = 0.08$, $p = .75$; or mixed methods, $b = -0.04$, $SE = 0.05$, $p = .46$.

3.3.3. Features of retell assessment

As shown in Table 3, oral or silent reading mode did not statistically moderate the association between retell and reading comprehension, $Q_m = 1.88$, $df = 1$, $p = .17$. The MES associated with studies using silent reading was not significantly larger than the MES associated with studies using oral reading mode, $b = 0.09$, $SE = 0.07$, $p = .17$.

It should be noted that we could not conduct moderation analysis for retell mode (i.e., oral versus written retell) because only two studies used solely written retell. To examine the moderating effect of number of prompts in retell assessment, we treated the number of prompts as a continuous variable when testing whether studies that used more prompts showed higher correlations (see Table 3). Again because only two studies used solely written retell, this moderation analysis was restricted to the studies that employed *oral* retell (n = 20, $k = 53$). Results revealed a significant positive association between the number of prompts and the magnitude of the association between oral retell and reading comprehension, $b = 0.11$, $SE = 0.04$, $p < .01$. Because the range of the number of prompts was one to three, the estimated correlation with reading comprehension for three prompts is .58 (0.25 intercept + 0.11*3).

We also examined the moderating effect of text genres in retell assessment. As Table 3 presents, text genre was not a significant moderator, $Q_m = 2.84$, $df = 2$, $p = .24$. Although the MES may appear larger for studies using narrative texts (which was the reference

group in the statistical model) compared with those using informational texts, $b = -.03$, $SE = 0.05$, $p = .63$, or both genres, $b = -0.18$, $SE = 0.11$, $p = .10$, the differences were not statistically significant.

We also sought to determine whether the magnitude of correlation between retell and other measures of reading comprehension is dependent on the way retell performance is evaluated. Given the limited number of studies involving written retell, we only examined whether *oral* retell evaluation methods (word counts, propositions or idea units, story structure elements, and overall quality) have a significant impact on effect sizes (n = 20, $k = 70$), and found that oral retell evaluation methods yielded no evidence of moderation effects, $Q_m = 0.73$, $df = 3$, $p = .87$ (see Table 3).

3.4. Research question 3: If there is a moderation effect of grade, is it explained by the other moderators (i.e., reading comprehension assessment method, retell assessment method, or retell evaluation method)?

As the grade level was found to significantly moderate the relation between retell and reading comprehension, we further explored whether the moderation by grade remains after controlling for the other moderators (see Table 4). As shown in Table 4, grade level was not statistically significant once the number of prompts or text genre of retell was respectively accounted for, indicating that the grade effect is largely explained by these factors. It is also notable that studies using cloze or maze format for reading comprehension assessment still produced stronger correlations between retell and reading comprehension than the studies using multiple-choice format, $b = 0.08$, $SE = 0.03$, $p < .05$ (Model 1), and the number of prompts still positively affected the correlation, $b = 0.11$, $SE = 0.04$, $p < .01$ (Model 3), even after controlling for grade level.

3.5. Research question 4: Does the effect of reading mode vary depending on grade level?

To test whether the effect of reading mode (oral or silent reading) on the relation varies by grade, an interaction between grade and reading mode was tested. Results in Table 4 show that the strength of the relation did not differ by grade, $b = -0.02$, $SE = 0.04$, $p = .54$ (Model 6).

4. Discussion

Retell is widely used as a measure of reading comprehension (Fuchs et al., 1988; Reed & Petscher, 2012; Shapiro et al., 2014). Our primary question in this study was the relation of retell to other measures of reading comprehension by quantitatively estimating the average effect size and systematically examining several potential moderators such as grade level, features of reading comprehension assessment method, and retell assessment features (reading mode, number of prompts, text genre, and evaluation method).

The present meta-analysis demonstrated that there is a moderate average correlation ($r = 0.46$) between retell and reading comprehension, corroborating a prior narrative review (Reed & Vaughn, 2012). Overall this finding indicates that students' retell does capture their reading comprehension skill. However, the moderate magnitude suggests that although retell can be one measure of evaluating the extent to which children comprehend written texts (reading comprehension), retell alone may not provide an entirely accurate picture of the student's reading comprehension.

Beyond the average estimate, there was large variation in the magnitudes of the relation between retell and reading comprehension (−0.26 to 0.95), and we systematically examined whether variation is explained by reader, text, and assessment factors. Several findings emerged. First, we found that the relation between retell and reading comprehension is slightly weaker in upper grades. This is in line with a previous study (Shapiro et al., 2014), and suggests that retell tends to be somewhat more sensitive to younger readers. Importantly, however, the weaker relation in upper grades disappeared once text genre or number of prompts in retell was accounted for, indicating that these factors explain the differences in the retell–reading comprehension relation as a function of grade. In other words, the weaker relations in upper grades are largely due to differences in text genres or number of prompts used in retell. The frequencies of effect sizes for narrative and informational texts in the included studies are presented by grade in Appendix E (a), which shows a trend that narrative texts were used in many studies with students in Grades 2 to 6. In contrast, informational texts and both types of texts were used less frequently across grades, but tended to be more frequently used in upper grades. These results show there is some overlap between grades and the type of text used, and this explains the differential relation between retell and other reading comprehension measures by grade. Furthermore, the moderation effect by grade also disappeared once the number of prompts was accounted for. As shown in Appendix E (b), most studies employed prompting twice while others ranged from one to three times. Importantly, no studies in Grade 6 and above employed prompting three times, indicating that the number of prompting was fewer in upper grades in extant studies, and this accounts for the weaker relation between retell and other reading comprehension measures in upper grades.

Our findings also underscore varying magnitudes of the relation between retell and reading comprehension as a function of features of reading comprehension assessment method. Specifically, the magnitudes for cloze or maze were stronger compared to those for multiple-choice format even after controlling for grade level. Reading comprehension is a complex construct where reader factors (working memory, inferencing skills, content knowledge, motivation) interact with text factors (demands on vocabulary, inference, content knowledge; Anderson & Armbruster, 1984; Kim, 2020; Matthew, 1996; McNeil, 1984; J. R. Miller & Kintsch, 1980; Rumelhart & Ortony, 1977; Schimmel & Ness, 2017; Snow, 2002). Studies have shown that reading comprehension tests are not comparable and interchangeable as various comprehension tasks measure multiple processes involved in comprehension to a different extent (Cutting & Scarborough, 2006; Keenan et al., 2008; Kendall, Mason, & Hunter, 1980; see the construction-integration model above). Explanations for the stronger relation of retell with reading comprehension measured by cloze or maze tasks are unclear and beyond the

scope of the present study. However, given the previous evidence that cloze and maze tasks do not draw on higher order inferential skills to a great extent (Cutting & Scarborough, 2006; Francis et al., 2005; Keenan et al., 2008; S. D. Miller & Smith, 1989), the stronger relation may indicate that retell may not readily tap into the higher order integration process that is necessary for deep comprehension (also see McNamara et al., 1996; but see Kida et al.'s (2016) argument that retell reveals readers' different levels of understanding including micro, macro, and superstructures). In fact, a recent study of Grade 4 students' retell of narrative and informational texts revealed that the vast majority of the retell was recall of the provided texts and only a limited number of utterances related to inferred information (Kim et al., 2020). If this is truly the case, then using retell as the sole measure of reading comprehension would not serve well in accurately capturing students' success or difficulties in deep or higher order comprehension skill. Future research is needed.

Finally, we found that the oral retell–reading comprehension relation was stronger with more prompts (the number of prompts administered in the reviewed studies ranged from 1 to 3), and this moderation effect remained after accounting for grade level. This finding is in line with a previous study (Reed & Petscher, 2012) and indicates that greater number of prompts (at least up to three prompts) may elicit more information about children's comprehension of the given material. However, this result should be taken with caution in a couple of aspects. First, it is not clear what the optimal frequency of prompts is—a simple increase in frequency asking for more information may not be effective. For example, asking children to add information several times after the child thought that s/he was done with her retell may elicit information that works against coherence of already presented information (i.e., the added information may seem tacked on and may render disorganized structure). Second, we only accounted for the number of prompts provided during oral retell assessments, and the effects of prompting conditions (when and how the prompts are provided) remain unclear as they were not always reported in studies. Future research is needed to examine (a) optimal ways of providing prompts, (b) the interaction between prompting condition and retell evaluation method (both oral and written retell), and (c) the effects of both aspects on the relation of retell and reading comprehension.

Several other potential moderators were also examined, including text genre of retell, reading mode in retell (oral reading or silent reading), and retell evaluation methods, but these factors did not make a difference in the strength of the relation between retell and reading comprehension. Note that we could not examine the moderation analysis for oral versus written retell due to a limited number of studies employing written retell. Although children tend to find narrative texts more comprehensible (Best et al., 2008; Graesser et al., 2003; Meyer & Ray, 2011; Mullis et al., 2003; Sáenz & Fuchs, 2002; Weaver & Kintsch, 1991; Williams et al., 2004), our study indicates the relation of retell and other measures of reading comprehension does not differ for narrative versus informational texts. An explanation for this finding is beyond the scope of this study as addressing this question adequately requires solid evidence about the comprehension processes that retell captures. If comprehension processes differ by genre, and retell captures comprehension processes fully, then the relation of retell and other measures of reading comprehension might differ. If retell does not fully capture multiple comprehension processes, then similar magnitudes of the relation for narrative and informational texts might reflect the limitation of retell in capturing multiple comprehension processes.

In addition, although a few studies found significant differences in reading comprehension performances between silent reading and oral reading (e.g., Hale et al., 2007; Prior & Welling, 2001), the average relation between retell and other measures of reading comprehension did not differ as a function of reading mode. If oral reading is particularly beneficial for children in the beginning phase of reading development or for those with reading difficulties (Fuchs et al., 1988; Kim et al., 2019), then the relation may differ by retell reading mode and grade level. However, results showed no difference in the strength of the relation by grade. Again note that this result does not indicate lack of a differential impact of oral versus silent reading on reading comprehension performance as a function of grade. Instead the results indicate no differential effect of reading mode by grade on the relation between retell and other reading comprehension measures, which would be impacted by comprehension processes that retell captures.

Finally, the results did not differ by different retell evaluation methods. Successful reading comprehension involves constructing accurate mental representations based on the given texts (Kintsch, 1988), and therefore, from a theoretical perspective, evaluating the extent of structural elements or overall quality may capture students' comprehension (or situation model) more accurately than counting the number of total retold words. Again, this question presumes that retell fully taps into multiple comprehension processes. If retell is limited in capturing deep comprehension (i.e., higher order integration), different retell evaluation methods may not make a difference in the relation of retell to other measures of reading comprehension. Given that studies with widely varying characteristics were included in the present study by the nature of meta-analysis, future studies are needed to address this question carefully.

5. Limitations

There are several limitations in the current study. First and foremost, the present results should be interpreted bearing in mind that the definition of retell and its operationalization (including instructions and prompting conditions) vary in research. In the present study, we employed the most common definition of retell where students were asked and/or purposely prompted to tell or write down the main ideas or everything they remembered from a passage they read, and included studies using this definition or operationalization. As shown in Table 2 and Fig. 2, respectively, there is large variation in retell measurement and in the relation of retell and other measures of reading comprehension. Although the results showed that on average, the retell defined in this manner is related to other measures of reading comprehension, the present study could not fully capture whether and how differences in retell measurement impacts the relation (other than the number of prompts).

Second, our meta-analysis was limited to the studies published in English with reported correlational coefficients between retell and reading comprehension. Consequently, potential articles that could provide valuable insight into the retell–reading comprehension correlation may have been excluded because they were published in languages other than English. Although some researchers have shown that there is no difference between language-restricted and language-inclusive meta-analyses (Moher et al., 2000;

Morrison et al., 2012), almost all the studies in our sample were conducted within the United States and their participants were not highly representative of the populations outside of the English-speaking world. Furthermore, the existence of the publication bias also calls for more unpublished studies to be included for a more precise estimation and higher generalizability.

Third, we could not examine moderation by other potential factors because of insufficient data such as English language learner (ELL) status, SES, gender differences, learning disability status, and interrater reliability on retell protocols. Although many studies provided descriptive information on these factors, few of them reported student performance by subgroup. Hence, we could not analyze data by these factors, and our results cannot inform any potential differences in the relation of retell to other measures of reading comprehension as a function of these factors. Reporting results by subgroups in studies would be extremely helpful for future efforts in investigating the moderation effects of these potential factors.

Furthermore, the studies that met our criteria contained a disproportionally small number of studies involving written retell measures, so the moderation analysis of retell mode could not be conducted and moderation analyses regarding prompts and retell evaluation methods were restricted to the studies using oral retell measures only. Additionally, relatively few studies included students in Grades 9 and above, which limited the developmental moderation analysis, and thus, the results should be interpreted with caution. Similarly, text genres in retell assessment was inconsistent across grades (see Appendix E), and therefore, future studies should strive for inclusion of both narrative and informational genres in retell assessments across grades.

It is also worth mentioning that retell and other reading comprehension in the vast majority of included studies were measured with less than perfect reliability. In fact, a substantial number of studies did not report reliability estimates for reading comprehension measures let alone validity information (see Appendix D). Measurement error attenuates the relation between retell and other measures of reading comprehension, and correcting for measurement error has been recommended (e.g., Hunter & Schmidt, 2004). However, there is ongoing debate about its appropriateness and issues on operational procedures (see Borenstein et al., 2011; Kepes, McDaniel, Brannick, & Banks, 2013). Therefore, we reported results using correlations reported in the studies without disattenuating them. However, when we conducted artifact corrections using the Psychmeta package (Version 2.3.6; Dahlke & Wiernik, 2019) in RStudio (Version 1.2.5033; RStudio Team, 2016), the overall average correlation was identical, 0.46.

Lastly, to maximize the effect sizes for data analysis, we employed pairwise deletion for all the regression models which resulted in the imbalanced sample sizes across models. Although the number of effect sizes included in this meta-analysis were large enough to estimate the overall effect, we were only allowed to assess each moderator with limited covariates (only 31 effect sizes would remain for running the potential full model where all the moderators could be taken into account simultaneously). Clearly, future investigation with the full regression model may provide more conclusive results and reliable implications.

6. Implications

The moderate relation of retell and other measures of reading comprehension cautions against using retell as the sole measure of reading comprehension. Retell should be accompanied with other measures, especially the ones tapping into higher order comprehension skills, to ensure a more valid and accountable interpretation of students' comprehension abilities. This might be particularly important in upper grades where the relation between retell and other reading comprehension measures is weaker. The present results brought up many questions and behoove future studies on retell. As noted above, reading comprehension is an active, interactive process, and therefore, it is important to investigate the extent to which retell captures various comprehension processes (what type of information is included in retell). Furthermore, a more systematic inquiry is needed on potential influence of retell assessment features, including modes of retell (oral or silent reading of texts; oral or written retell) and instructions for retell (e.g., number and nature of prompts), and their potential interactions with each other, with reader characteristics (e.g., beginning or struggling readers; inferencing ability), and with text characteristics (e.g., inferential demands).

7. Conclusion

Our investigation provides evidence for a moderate relation of retell to reading comprehension. This meta-analysis serves as a comprehensive quantitative synthesis of the relation between retell and reading comprehension, which contributes to the growing literature on the validity of retell instruments, and advances our understanding of the strength and limitations of retell as a measure of reading comprehension for students across grades.

Acknowledgements

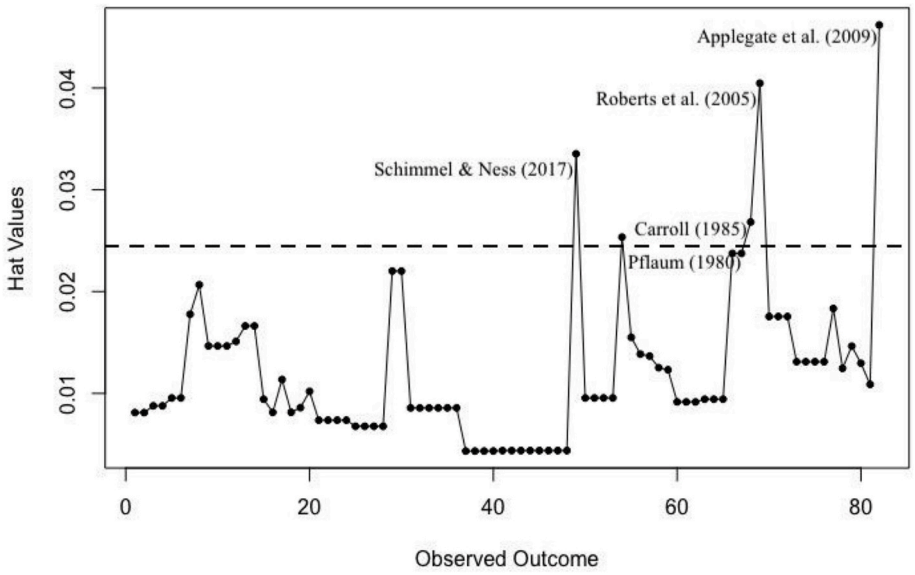
This research was partly supported by grants from the Institute of Education Sciences, US Department of Education (R305A180055) and the Eunica Kennedy Shriver National Institute of Child Health and Human Development (NICHD; P50HD052120) to the second author. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agency.

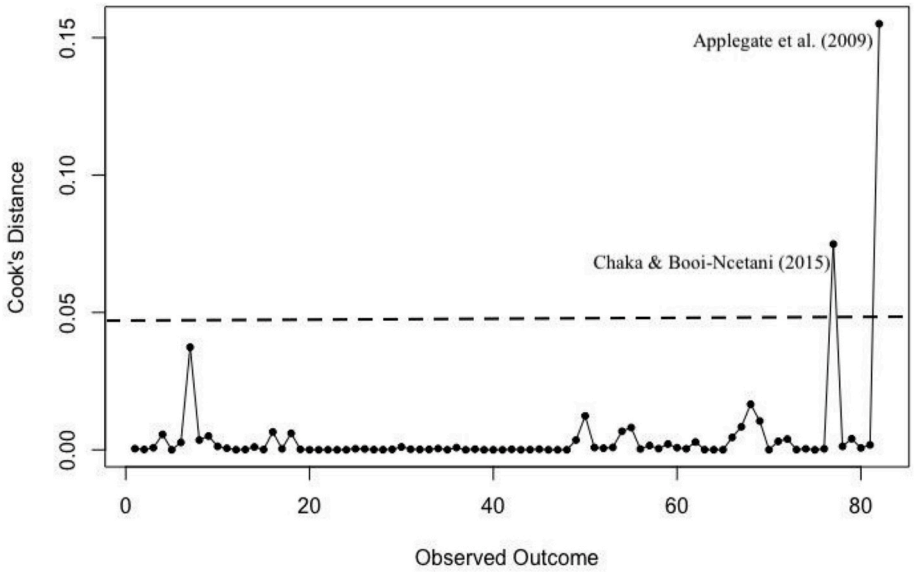
Appendix F. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.edurev.2020.100375.

Appendix A

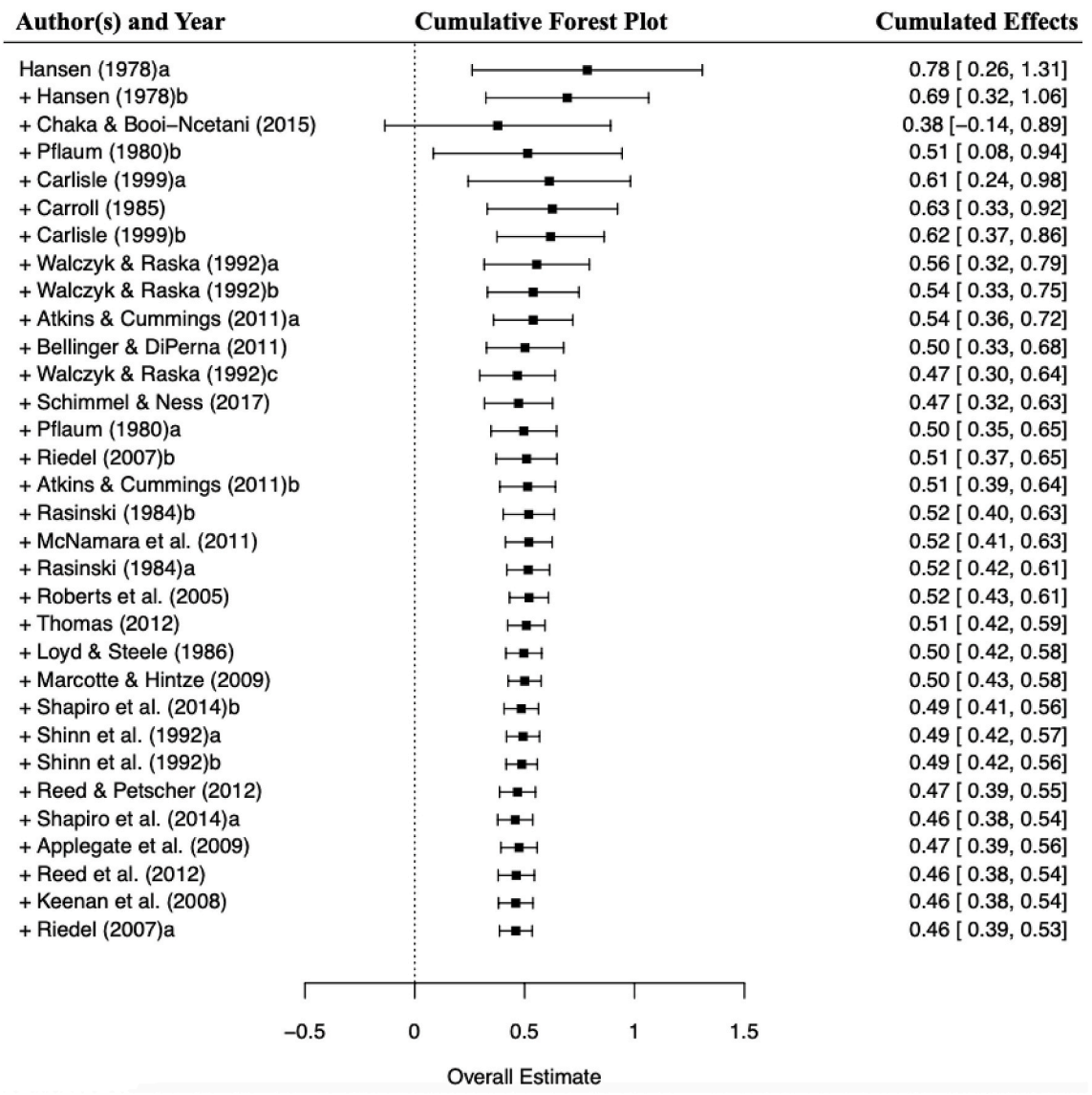
Influence analysis plots





Appendix B

Cumulative Forest Plot



Appendix C

Study Quality Assessment

1. Was the research question or objective clearly stated?
2. Was the study population clearly defined?
3. Was the sample selected from the same or similar populations and clearly described (i.e., age, gender, race, disabilities)?
4. Were inclusion and exclusion criteria for being in the study pre-specified and applied uniformly to all participants?
5. Was sample size justification, power description, or variance and effect estimates provided?
6. For longitudinal studies, was attrition (withdrawals and drop-outs) reported in terms of numbers and/or reasons per group?
7. Were the reading assessments (both retell and reading comprehension tests) clearly defined across all study participants?
8. Were the reading assessments (both retell and reading comprehension tests) reliable across all study participants?
9. Were the reading assessments (both retell and reading comprehension tests) implemented consistently across all study participants?
10. Did they report basic information (i.e., mean, standard deviations, range, correlations for all assessments given)?
11. Did they report reliability estimates on all measures? And, if so, were they above 0.7?

12. Were the statistical measures appropriate for the study design?
Yes (Strong) = 0 Somewhat (Moderate) = 1 No (weak) = 2

Appendix D

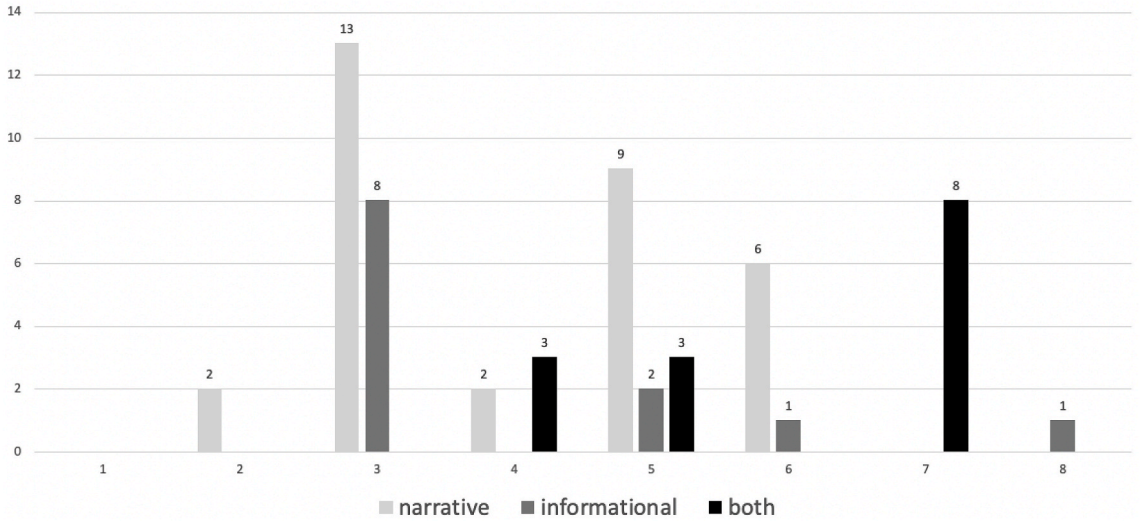
Reading comprehension measures and reported reliability/validity information.

Study author(s)	Test name	Reliability	Validity
Applegate et al. (2009)	Critical Reading Inventory-2	N/A	N/A
Atkins and Cummings (2011)	ITBS	N/A	.64
	MontCAS	N/A	N/A
	RTF	.57	.51
Bellinger and Diperna (2011)	WJ-PC	.98	.71
	RTF	72%	.59
Carlisle (1999)	Sentence Verification tests	N/A	.56
	Free recall task	98%	N/A
Carroll (1985)	Metropolitan Achievement Test	N/A	N/A
	Retelling	[.86,.98]	N/A
Chaka and Booi-Ncetani (2015)	Reading comprehension test	N/A	N/A
	Reading recall task	N/A	N/A
Hansen (1978)	Comprehension test	N/A	N/A
	Retell test	94.1%	N/A
Keenan et al. (2008)	QRI	N/A	N/A
	PIAT	N/A	N/A
	WJPC	N/A	N/A
	GORT	N/A	N/A
Loyd and Steele (1986)	SRA	.91	N/A
	Written recall	.81	N/A
Marcotte and Hintze (2009)	GRADE	[.69,.83]	>.90
	Sentence verification technique	[.70,.80]	.73
	Maze	.83	N/A
	RTF	.59	N/A
	Written retell	.97	[.76,.81]
McNamara et al. (2011)	Multiple-choice questions	N/A	N/A
	Recall tasks	>90%	N/A
Pflaum (1980)	ITBS	N/A	N/A
	CTBS	N/A	N/A
	Retelling comprehension	>85%	N/A
Rasinski (1984)	Gates-MacGinitie Reading Test	100%	N/A
	Multiple choice comprehension test	100%	N/A
	Retelling	.94	N/A
Reed and Petscher (2012)	TAKS	N/A	N/A
	TMSFA-Retell	.978	N/A
Reed et al. (2012)	AIMSweb Reading Maze	.81	N/A
	GRADE	.90	N/A
	WJPC	.88	N/A
	TAKS	high .80s to low .90s	N/A
	TMSFA-Retell	63%	N/A
Riedel (2007)	GRADE	.94	.68
	TerraNova	.89	.68
	RTF	N/A	N/A
Roberts et al. (2005)	WDRB	.95	N/A
	VIP retell fluency	.73	N/A
Schimmel and Ness (2017)	Qualitative Reading Inventory-5	.98	.71
Shapiro et al. (2014)	PSSA	.90	N/A
	Reading Retell Rubric	86%	N/A
	Adapted RTF	89%	N/A
Shinn et al. (1992)	SDRT	.80	N/A
	Cloze	.91	[.47,.73]
	Written retell	.99	[.60,.79]
Thomas (2012)	GRADE	.94	.86
	PSSA	N/A	N/A
	4Sight	.73	[.81,.89]
	Reading Retell Rubric	92.5%	N/A
	Adapted RTF	90.5%	[.39,.81]
Walczyk and Raska (1992)	ITBS	N/A	N/A
	Reading inference test	N/A	N/A
	Recall inference test	86%	N/A

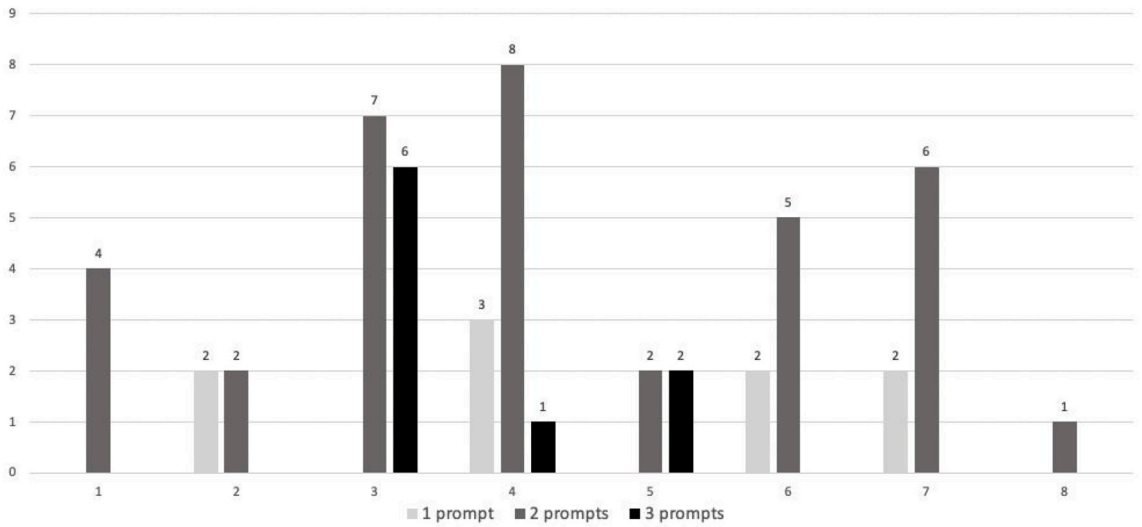
Note. For studies that reported multiple reliabilities (e.g., by grade or passage), the mean reliability was calculated for presentation. For studies that did not report validity but provided the correlations with other reading assessments, the mean correlation was calculated and presented as the validity coefficient. N/A = not available (primary studies did not report information). ITBS = Iowa Tests of Basic Skills; MontCAS = Montana State Criterion-Referenced Test; RTF = Dynamic Indicators of Basic Early Literacy Skills-Retell Fluency; WJ-PC = The Woodcock–Johnson Passage Comprehension and Reading Vocabulary; QRI = Qualitative Reading Inventory; PIAT = Peabody Individual Achievement Test; WJPC = Woodcock–Johnson Passage Comprehension subtest; GORT = Gray Oral Reading Test; SRA = Science Research Associates Achievement; GRADE = Group Reading Assessment and Diagnostic Evaluation; CTBS = Comprehensive Test of Basic Skills; TAKS = Texas Assessment of Academic Skills; TMSFA = Texas Middle School Fluency Assessment; WDRB = Woodcock Diagnostic Reading Battery; VIP = Vitals Indicators of Progress system; PSSA = Pennsylvania System of School Assessment; SDRT = Stanford Diagnostic Reading Test; 4Sight = Pennsylvania 4Sight Reading Benchmark Assessment.

Appendix E

(a) The frequency of effect sizes for narrative texts and informational texts by grade (X axis).



(b) The frequency of effect sizes as a function of the number of prompts by grade (X axis).



References

References marked with an asterisk indicate studies included in the meta-analysis.

Adlof, S. M., Catts, H. W., & Little, T. D. (2006). Should the simple view of reading include a fluency component? *Reading and Writing, 19*(9), 933–958. https://doi.org/10.1007/s11145-006-9024-z

Alderson, J. C. (1978). Cloze procedures. In O. K. Buros (Ed.), *The eighth mental measurements yearbook* (pp. 1171–1174). Highland Park: NJ: Gryphon Press.

Allington, R., & McGill-Franzen, A. (2010). Why so much oral reading? In E. H. Hiebert, & D. R. Reutzel (Eds.), *Revisiting silent reading: New directions for teachers and researchers* (pp. 45–56). https://doi.org/10.1598/0833.03

Anderson, T. H., & Armbruster, B. B. (1984). Studying. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 657–680). White Plains: NY: Longman.

Andreassen, R., & Bråten, I. (2009). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading, 33*(3), 263–283. https://doi.org/10.1111/j.1467-9817.2009.01413.x

* Applegate, M. D., Applegate, A. J., & Modla, V. B. (2009). "She's my best reader; she just can't comprehend": Studying the relationship between fluency and comprehension. *The Reading Teacher, 62*(6), 512–521. https://doi.org/10.1598/rt.62.6.5.

* Atkins, T. L., & Cummings, K. D. (2011). Utility of oral reading and retell fluency in predicting proficiency on the Montana comprehensive assessment system. *Rural Special Education Quarterly, 30*(2), 3–12. https://doi.org/10.1177/875687051103000202.

Barnes, A. E., Kim, Y. S., & Phillips, B. M. (2014). The relations of proper character introduction to narrative quality and listening comprehension for young children from high poverty schools. *Reading and Writing, 27*(7), 1189–1205. https://doi.org/10.1007/s11145-013-9481-0

Barth, A. E., Barnes, M., Francis, D. J., Vaughn, S., & York, M. (2015). Inferential processing among adequate and struggling adolescent comprehenders and relations to reading comprehension. *Reading and Writing, 28*, 587–609. https://doi.org/10.1007/s11145-014-9540-1

* Bellinger, J. M., & Diperna, J. C. (2011). Is fluency-based story retell a good indicator of reading comprehension?. *Psychology in the Schools, 48*(4), 416–426. https://doi.org/10.1002/pits.20563.

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology, 29*(2), 137–164. https://doi.org/10.1080/02702710801963951

Blachowicz, C. L. Z., & Ogle, D. (2008). *Reading comprehension ;: Strategies for independent learners*. Guilford Press.

Bontis, N., Hardie, T., & Serenko, A. (2009). Techniques for assessing skills and knowledge in a business strategy classroom. *International Journal of Teaching and Case Studies, 2*(2), 162. https://doi.org/10.1504/ijtcs.2009.031060

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

van den Broek, P., Rapp, D. N., & Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Processes, 39*(2–3), 299–316. https://doi.org/10.1080/0163853x.2005.9651685

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing, 11*, 489–503. https://doi.org/10.1023/A:1008084120205

Cain, K., & Oakhill, J. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology, 76*(4), 697–708. https://doi.org/10.1348/000709905X69807

* Carlisle, J. F. (1999). Free recall as a test of reading comprehension for students with learning disabilities. *Learning Disability Quarterly, 22*(1), 11–22. https://doi.org/10.2307/1511148.

* Carroll, R. G. (1985). *The use of oral retellings of basal stories*. University of Florida.

* Chaka, C., & Booi-Ncetani, N. N. (2015). An investigation into the English reading comprehension of Grade 10 English first additional language learners at a senior secondary school. *Reading and Writing, 6*(1), 1–7. https://doi.org/10.4102/rw.v6i1.62.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Collins, A. A., Compton, D. L., Lindström, E. R., & Gilbert, J. K. (2020). Performance variations across reading comprehension assessments: Examining the unique contributions of text, activity, and reader. *Reading and Writing, 33*(3), 605–634. https://doi.org/10.1007/s11145-019-09972-5

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*(3), 277–299. https://doi.org/10.1207/s1532799xssr1003_5

Dahlke, J. A., & Wiernik, B. M. (2019). Psychmeta: An R package for psychometric meta-analysis. *Applied Psychological Measurement, 43*(5), 415–416. https://doi.org/10.1177/0146621618795933

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader–text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*(3), 515–528. https://doi.org/10.1037/a0027182

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). Package ' robumeta'. Retrieved from http://cran.uni-muenster.de/web/packages/robumeta/robumeta.pdf.

Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies? *Educational Psychology Review, 23*(4), 553–576. https://doi.org/10.1007/s10648-011-9175-6

Follmer, D. J., & Sperling, R. A. (2018). Interactions between reader and text: Contributions of cognitive processes, strategy use, and text cohesion to comprehension of expository science text. *Learning and Individual Differences, 67*, 177–187. https://doi.org/10.1016/j.lindif.2018.08.005

Fox, E., & Alexander, P. (2017). Text comprehension: A retrospective, perspective, and prospective. In S. E. Israel, & G. G. Duffy (Eds.), *Handbook of research on reading comprehension, second edition* (pp. 335–352). New York: Guilford Press.

Francis, D. J., Fletcher, J. M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. In S. G. Paris, & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 369–394). Mahwah: NJ: Erlbaum.

Francis, D. J., Kulesz, P. A., & Benoit, J. S. (2018). Extending the simple view of reading to account for variation within readers and across texts: The Complete view of reading (CVR i). *Remedial and Special Education, 39*(5), 274–288. https://doi.org/10.1177/0741932518772904

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*(1), 45–58. https://doi.org/10.1080/02796015.1992.12085594

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20–28. https://doi.org/10.1177/074193258800900206

García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research, 84*(1), 74–111. https://doi.org/10.3102/0034654313499616

Good, R. H., III, & Kaminski, R. A. (2002). Dynamic indicators of basic early literacy skills TM 6 th edition DIBELS TM preschool kindergarten first grade second grade third grade Administration and scoring guide dynamic indicators of basic early literacy skills (DIBELS) 6th edition edited by. Retrieved from http://dibels.uoregon.edu/.DIBELSTM.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 6–10. https://doi.org/10.1177/074193258600700104

Graesser, A. C., León, J. A., & Otero, J. (2002). Introduction to the psychology of science text comprehension. In J. Otero, J. A. León, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 1–15). Mahwah, NJ: Erlbaum.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet, & C. E. Snow (Eds.), *Rethinking reading comprehension* (p. 82). New York: Guilford Press.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371–395. https://doi.org/10.1037/0033-295x.101.3.371

Hale, A. D., Hawkins, R. O., Sheeley, W., Reynolds, J. R., Jenkins, S., Schmitt, A. J., et al. (2011). An investigation of silent versus aloud reading comprehension of elementary students using Maze assessment procedures. *Psychology in the Schools, 48*(1), 4–13. https://doi.org/10.1002/pits.20543

Hale, A. D., Skinner, C. H., Williams, J., Hawkins, R., Neddenriep, C. E., & Dizer, J. (2007). Comparing comprehension following silent and aloud reading across elementary and secondary students: Implication for curriculum-based measurement. *Behavior Analyst Today, 8*(1), 9–23. https://doi.org/10.1037/h0100101
Hall, S., Basran, J., Paterson, K. B., Kowalski, R., Filik, R., & Maltby, J. (2014). Individual differences in the effectiveness of text cohesion for science text comprehension. *Learning and Individual Differences, 29*, 74–80. https://doi.org/10.1016/j.lindif.2013.10.014
* Hansen, C. L. (1978). Story retelling used with average and learning disabled readers as a measure of reading comprehension. *Learning Disability Quarterly, 1*(3), 62–69. https://doi.org/10.2307/1510938.
Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39–65. https://doi.org/10.1002/jrsm.5
Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557
Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127–160. https://doi.org/10.1007/BF00401799
Howell, K., & Nolet, V. (2000). *Curriculum-based evaluation: Teaching and decision-making, 3rd ed.* Scarborough, Ontario: Wadsworth Thomson Learning.
Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.
Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59*(5), 421–432. https://doi.org/10.1177/001440299305900505
Juel, C., & Holmes, B. (1981). Oral and silent reading of sentences. *Reading Research Quarterly, 16*(4), 545–568. https://doi.org/10.2307/747315
* Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*(3), 281–300. https://doi.org/10.1080/10888430802132279.
Kendall, J. R., Mason, J. M., & Hunter, W. (1980). Which comprehension? Artifacts in the measurement of reading comprehension. *Journal of Educational Research, 73*(4), 233–236. https://doi.org/10.1080/00220671.1980.10885242
Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. (2013). Meta-analytic reviews in the organizational Sciences: Two meta-analytic schools on the way to MARS (the meta-analytic reporting standards). *Journal of Business and Psychology, 28*(2), 123–143. https://doi.org/10.1007/s10869-013-9300-2
Kida, A., de, S. B., Ávila, C. R. B. de, & Capellini, S. A. (2016). Reading comprehension assessment through retelling: Performance profiles of children with dyslexia and language-based learning disability. *Frontiers in Psychology, 7*, 787. https://doi.org/10.3389/fpsyg.2016.00787
Kim, Y.-S. G. (2020). Hierarchical and dynamic relations of language and cognitive skills to reading comprehension: Testing the direct and indirect effects model of reading (DIER). *Journal of Educational Psychology, 112*(4), 667–684. https://doi.org/10.1037/edu0000407
Kim, Y.-S. G., Dore, R. A., Cho, M., Golinkoff, R., & Amendum, S. (2020). *Theory of mind, mental state talk, and discourse comprehension: Theory of mind process is more important for narrative comprehension than for informational text comprehension*. Manuscript submitted for publication.
Kim, Y.-S. G., & Petscher, Y. (2020). Influences of person, text, and assessment factors on text/discourse comprehension in oral language (listening comprehension). *Annals of Dyslexia*. https://doi.org/10.1007/s11881-020-00208-8
Kim, Y. S. G., Petscher, Y., & Vorstius, C. (2019). Unpacking eye movements during oral and silent reading and their relations to reading proficiency in beginning readers. *Contemporary Educational Psychology, 58*, 102–120. https://doi.org/10.1016/j.cedpsych.2019.03.002
Kim, Y. S. G., & Wagner, R. K. (2015). Text (oral) reading fluency as a construct in reading development: An investigation of its mediating role for children from grades 1 to 4. *Scientific Studies of Reading, 19*(3), 224–242. https://doi.org/10.1080/10888438.2015.1007375
Kim, Y.-S. G., & Schatschneider, C. (2017). Expanding the developmental models of writing: A direct and indirect effects model of developmental writing (DIEW). *Journal of Educational Psychology, 109*, 35–50. https://doi.org/10.1037/edu0000129
King, D. J. (1960). On the accuracy of written recall; a scaling and factor analytic study. *Psychological Record, 10*(2), 113–122. https://doi.org/10.1007/bf03393351
Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, New Jersey: Lawrence Erlbaum Association.
Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163
Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363
Kragler, S. (1995). The transition from oral to silent reading. *Reading Psychology, 16*(4), 395–408. https://doi.org/10.1080/0270271950160402
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE publications, Inc.
* Loyd, B. H., & Steele, J. L. (1986). Assessment of reading comprehension: A comparison of constructs. *Reading Psychology, 7*(1), 1–10. https://doi.org/10.1080/0270271860070102.
* Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology, 47*(5), 315–335. https://doi.org/10.1016/j.jsp.2009.04.003.
Matthew, K. I. (1996). The impact of CD-ROM storybooks on children's reading comprehension and reading attitude. *Journal of Educational Multimedia and Hypermedia, 5*(3), 379–394. https://doi.org/10.5555/251735.251744
Mazany, T., Pimentel, S., Orr, C. S., & Crovo, M. (2015). Reading framework for the 2015 national assessment of educational progress national assessment governing board. *Assessment*. Retrieved from https://www.nagb.org/content/nagb/assets/documents/publications/frameworks/reading/2015-reading-framework.pdf.
McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning form text. *Cognition and Instruction, 14*(1), 1–43.
McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model for comprehension. In B. Ross (Ed.), *The psychology of learning and motivation* (pp. 297–384). New York, NY: Elsevier.
* McNamara, D. S., Ozuru, Y., & Floyd, R. G. (2011). Comprehension challenges in the fourth grade: The roles of text cohesion, text genre, and readers' prior knowledge. *International Electronic Journal of Environmental Education, 4*(1), 229–257. Retrieved from https://www.iejee.com/index.php/IEJEE/article/view/222.
McNamara, D. S., & Shapiro, A. M. (2005). Multimedia and hypermedia solutions for promoting metacognitive engagement, coherence, and learning. *Journal of Educational Computing Research, 33*(1), 1–29. https://doi.org/10.2190/7N6R-PCJL-UMHK-RYPJ
McNeil, J. D. (1984). *New directions for classroom practice. Scott, Foresman*. Reading comprehension.
Meyer, B. J., & McConkie, G. W. (1973). What is recalled after hearing a passage? *Journal of Educational Psychology, 65*(1), 109–117. https://doi.org/10.1037/h0034762
Meyer, B. J. F., & Ray, M. N. (2011). Structure strategy interventions: Increasing reading comprehension of expository text. *International Electronic Journal of Environmental Education, 4*(1), 127–152. Retrieved from https://www.iejee.com/index.php/IEJEE/article/view/217.
Miller, J. R., & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning & Memory, 6*(4), 335. https://doi.org/10.1037/0278-7393.6.4.335
Miller, S. D., & Smith, D. E. P. (1985). Differences in literal and inferential comprehension after reading orally and silently. *Journal of Educational Psychology, 77*(3), 341–348. https://doi.org/10.1037/0022-0663.77.3.341
Miller, S. D., & Smith, D. E. P. (1989). Relations among oral reading, silent reading and listening comprehension of students at differing competency levels. *Reading Research & Instruction, 29*(2), 73–84. https://doi.org/10.1080/19388079009558006
Mills, C. B., Diehl, V. A., Birkmire, D. P., & Mou, L.-C. (1993). Procedural text: Predictions of importance ratings and recall by models of reading comprehension. *Discourse Processes, 16*(3), 279–315. https://doi.org/10.1080/01638539309544841
Moher, D., Pham, B., Klassen, T. P., Schulz, K. F., Berlin, J. A., Jadad, A. R., et al. (2000). What contributions do languages other than English make on the results of meta-analyses? *Journal of Clinical Epidemiology, 53*(9), 964–972. https://doi.org/10.1016/s0895-4356(00)00188-8
Morrison, A., Polisena, J., Husereau, D., Moulton, K., Clark, M., Fiander, M., … Rabb, D. (2012). The effect of English-language restriction on systematic review-based meta-analyses: A systematic review of empirical studies. *International Journal of Technology Assessment in Health Care, 28*(2), 138–144. https://doi.org/10.1017/S0266462312000086

Morrow, L. M. (1985). Retelling stories: A strategy for improving young children's comprehension, concept of story structure, and oral language complexity. *The Elementary School Journal, 85*(5), 647–661. https://doi.org/10.1086/461427
Morrow, L. M. (1986). Effects of structural guidance in story retelling on children's dictation of original stories. *Journal of Reading Behavior, 18*(2), 135–152. https://doi.org/10.1080/10862968609547561
Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). PIRLS 2001 international report IEA's study of reading literacy achievement in primary school in 35 countries PIRLS. Retrieved from http://pirls.bc.edu.
Nakagawa, S., & Santos, E. S. A. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology, 26*(5), 1253–1274. https://doi.org/10.1007/s10682-012-9555-5
National Heart, Lung, and B. I. Study quality assessment tools | national Heart, Lung, and Blood Institute (NHLBI). n.d, Retrieved from https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools.
National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4769)*. Washington, DC: U.S.
Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67*(3), 359–370. https://doi.org/10.1111/j.2044-8279.1997.tb01250.x
Parker, R., Hasbrouck, J. E., & Tindal, G. (1992). The maze as a classroom-based reading measure: Construction methods, reliability, and validity. *The Journal of Special Education, 26*(2), 195–218. https://doi.org/10.1177/002246699202600205
Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687
* Pflaum, S. W. (1980). The predictability of oral reading behaviors on comprehension in learning disabled and normal readers. *Journal of Literacy Research, 12*(3), 231–236. https://doi.org/10.1080/10862968009547374.
Prior, S. M., & Welling, K. A. (2001). "Read in your head": A vygotskian analysis of the transition from oral to silent reading. *Reading Psychology, 22*(1), 1–15. https://doi.org/10.1080/02702710151130172
R Core Team. (2018). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org/.
RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA.
* Rasinski, T. V. (1984). *Developing models of reading fluency*. Newark, DE: International Reading Association.
Rasinski, T. V., Rikli, A., & Johnston, S. (2009). Reading fluency: More than automaticity? More than a concern for the primary grades? *Literacy Research and Instruction, 48*(4), 350–361. https://doi.org/10.1080/19388070802468715
* Reed, D. K., & Petscher, Y. (2012). The influence of testing prompt and condition on middle school students' retell performance. *Reading Psychology, 33*(6), 562–585. https://doi.org/10.1080/02702711.2011.557333.
Reed, D. K., & Vaughn, S. (2012). Retell as an indicator of reading comprehension. *Scientific Studies of Reading, 16*(3), 187–217. https://doi.org/10.1080/10888438.2010.538780
* Reed, D. K., Vaughn, S., & Petscher, Y. (2012). The validity of a holistically scored retell protocol for determining the reading comprehension of middle school students. *Learning Disability Quarterly, 35*(2), 76–89. https://doi.org/10.1177/0731948711432509.
* Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*(4), 546–567. https://doi.org/10.1598/rrq.42.4.5.
* Roberts, G., Good, R., & Corcoran, S. (2005). Story retell: A fluency-based indicator of reading comprehension. *School Psychology Quarterly, 20*(3), 304–317. https://doi.org/10.1521/scpq.2005.20.3.304.
RStudio Team. (2016). RStudio: Integrated development for R. Retrieved from http://www.rstudio.com/.
Rumelhart, D. E., & Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, & W. E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 99–135). Retrieved from http://www.sciepub.com/reference/69865.
Sadoski, M., Goetz, E. T., & Fritz, J. B. (1993). A causal model of sentence recall: Effects of familiarity, concreteness, comprehensibility, and interestingness. *Journal of Reading Behavior, 25*(1), 5–16. https://doi.org/10.1080/10862969309547799
Sáenz, L. M., & Fuchs, L. S. (2002). Examining the reading difficulty of secondary students with learning disabilities. *Remedial and Special Education, 23*(1), 31–41. https://doi.org/10.1177/074193250202300105
Salvia, J., & Hughes, C. (1990). *Curriculum-based assessment: Testing what is taught*. New York: MacMillan Publishing Company.
* Schimmel, N., & Ness, M. (2017). The effects of oral and silent reading on reading comprehension. *Reading Psychology, 38*(4), 390–416. https://doi.org/10.1080/02702711.2016.1278416.
Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and Education, 12*(4), 431–459. https://doi.org/10.1016/S0898-5898(01)00073-0
* Shapiro, E. S., Fritschmann, N. S., Thomas, L. B., Hughes, C. L., & McDougal, J. (2014). Concurrent and predictive validity of reading retell as a brief measure of reading comprehension for narrative text. *Reading Psychology, 35*(7), 644–665. https://doi.org/10.1080/02702711.2013.790328.
* Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education, 34*(3), 164–172. https://doi.org/10.1177/002246690003400305.
Shinn, M. R., Good, Knutson, N., Tilly, & Collins, V. I. L. (1992). Oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*(3), 459–479.
Singer, M., Graesser, A. C., & Trabasso, T. (1994). Minimal or global inference during reading. *Journal of Memory and Language, 33*(4), 421–441. https://doi.org/10.1006/jmla.1994.1020
Smith, E. E., & Swinney, D. A. (1992). The role of schemas in reading text: A real-time examination. *Discourse Processes, 15*(3), 303–316. https://doi.org/10.1080/01638539209544814
Snow, C. (2002). Reading for understanding: Toward an R&D program in reading comprehension. RAND corporation.
Stein, N. L., & Glenn, C. G. (1975). *An analysis of story comprehension in elementary school children: A test of a schema. ERIC document reproduction service, ED 121 474*.
Sterne, J. A. C., & Egger, M. (2006). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). https://doi.org/10.1002/0470870168.ch6
Suhorsky, J. (1975). *An investigation of the relationship between undeleted text preceding a cloze test and cloze test results*. College Park: University of Maryland.
Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology, 2*(1), 85–112. https://doi.org/10.1007/s40865-016-0026-5
* Thomas, L. B. (2013). Evaluating a brief measure of reading comprehension for narrative and expository text: The convergent and predictive validity of the reading retell rubric. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 73*, 10-A(E).
Tilstra, J., & McMaster, K. L. (2013). Cognitive processes of middle grade readers when reading expository text with an assigned goal. *Learning and Individual Differences, 28*, 66–74. https://doi.org/10.1016/j.lindif.2013.09.011
Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics, 40*(6), 604–634. https://doi.org/10.3102/1076998615606099
Trabasso, T., & Magliano, J. P. (1996). Conscious understanding during comprehension. *Discourse Processes, 21*, 255–287.
Veritas Health Innovation. (2018). Covidence systematic review software. Retrieved from www.covidence.org.
Vevea, J. L., & Coburn, K. M. (2015). Maximum-likelihood methods for meta-analysis: A tutorial using R. *Group Processes & Intergroup Relations, 18*(3), 329–347. https://doi.org/10.1177/1368430214558311
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03
Vieiro, P., & García-Madruga, J. A. (1997). An analysis of story comprehension through spoken and written summaries in school-age children. *Reading and Writing, 9*(1), 41–53. https://doi.org/10.1023/A:1007932429184

* Walczyk, J. J., & Raska, L. J. (1992). The relation between low- and high-level reading skills in children. *Contemporary Educational Psychology, 17*(1), 38–46. https://doi.org/10.1016/0361-476X(92)90044-Y.

Weaver, C., & Kintsch, W. (1991). Expository text. *The Handbook of Reading Research, 2*, 230–245.

Williams, J. P., Hall, K. M., & Lauer, K. D. (2004). Teaching expository text structure to young at-risk learners: Building the basics of comprehension instruction. *Exceptionality, 12*(3), 129–144. https://doi.org/10.1207/s15327035ex1203_2

Wolfe, M. B. W., & Mienko, J. A. (2007). Learning and memory of factual content from narrative and expository text. *British Journal of Educational Psychology, 77*(3), 541–564. https://doi.org/10.1348/000709906X143902

Yuill, N., & Oakhill, J. (1991). *Children's problems in text comprehension: An experimental investigation*. Cambridge, United Kingdom: Cambridge University Press.